

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen



The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

ABSTRACT

THE DEVELOPMENT AND EVALUATION OF SOFTWARE TO FOSTER PROFESSIONAL DEVELOPMENT IN EDUCATIONAL ASSESSMENT

**by
Morgan C. Benton**

This dissertation sought to answer the question: Is it possible to build a software tool that will allow teachers to write better multiple-choice questions? The thesis proceeded from the finding that the quality of teaching is very influential in the amount that students learn. A basic premise of this research, then, is that improving teachers will improve learning. With this foundation, the next question became what area of teaching to improve. The literature on educational assessment indicated that teachers lack competence at effective assessment, particularly in the area of multiple-choice question generation. It is likely that improvement in this area would yield large gains in educational achievement by students.

Several areas of literature including teacher professional development, modification of health-related behaviors, and the information systems theories of captology and structuration theory were synthesized to develop a general model for designing systems to foster teacher professional development. This model was then applied to design and build a tool, QuesGen—a web-based system to help teachers write better multiple-choice questions. The tool was evaluated. Quantitative and qualitative results are presented, their implications discussed, and future steps are laid out.

**THE DEVELOPMENT AND EVALUATION OF SOFTWARE TO FOSTER
PROFESSIONAL DEVELOPMENT IN EDUCATIONAL ASSESSMENT**

**by
Morgan C. Benton**

**A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Information Systems**

Department of Information Systems

January 2008

Copyright © 2008 by Morgan C. Benton

ALL RIGHTS RESERVED

APPROVAL PAGE

**THE DEVELOPMENT AND EVALUATION OF SOFTWARE TO FOSTER
PROFESSIONAL DEVELOPMENT IN EDUCATIONAL ASSESSMENT**

Morgan C. Benton

5/07/2007

Dr. Marilyn M. Tremaine, Dissertation Advisor
Professor Emerita of Information Systems, NJIT

Date

5/7/07

Dr. Starr Roxanne Hiltz, Committee Member
Distinguished Professor of Information Systems, NJIT

Date

5/7/07

Dr. Michael Bieber, Committee Member
Professor of Information Systems, NJIT

Date

05/07/07

Dr. Katia Passerini, Committee Member
Assistant Professor of Management Information Systems, NJIT

Date

5/7/07

Dr. Rachelle S. Heller, Committee Member
Professor of Computer Science
Associate Dean for Academic Affairs, Mount Vernon Campus
Director, Elizabeth Somers Women's Leader Program
George Washington University

Date

BIOGRAPHICAL SKETCH

Author: Morgan C. Benton
Degree: Doctor of Philosophy
Major: Information Systems
Date: January 2008

Education:

- Doctor of Philosophy in Information Systems
New Jersey Institute of Technology, Newark, NJ, 2008
- Master of Science in Information Systems
New Jersey Institute of Technology, Newark, NJ, 2002
- Bachelor of Arts in Leadership Studies and Sociology
University of Richmond, VA, 1996

This is dedicated to Nozomi, Athena, and Molly for your enormous patience while I finished this work, to my mother and father who are responsible for my deep and abiding concern for teaching and learning, and to students everywhere who are frustrated and confused by poorly-written multiple-choice questions (some by me).

ACKNOWLEDGMENT

Though only a precious few will ever read these words, I know that my good friend and invaluable thesis advisor, Dr. Marilyn M. Tremaine, would not wish me to spend more words on her than a simple thank you. Thank you, Marilyn.

Thank you also to the members of my dissertation committee—Dr. Starr Roxanne Hiltz, Dr. Michael Bieber, Dr. Katia Passerini, and Dr. Rachelle Heller. Your guidance kept me on a productive path and allowed me to reach a successful conclusion.

I would also like to thank Drs. Starr Roxanne Hiltz and Murray Turoff whom I met as professors online while working towards a master's degree in IS via distance learning in Japan. Murray and Roxanne not only encouraged me to pursue the Ph.D., but made it possible by securing a lectureship for me at NJIT which allowed me to support my family for the duration of my study. For that, and for a great many other kindnesses, I and my family are truly grateful.

Finally, I would like to thank several family members, friends and colleagues, without whose support this endeavor would have been *very* much more difficult: Claire Benton Tunkel, Gene Moon, Michele Collins, Hyo-Joo Han, Sukeshini Grandhi, Vivek Vadavattath.

TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION.....	1
1.1 Motivation	1
1.2 Research Questions	2
1.3 Organization of the Dissertation.....	6
2 THE STATE OF THE ART.....	8
2.1 Introduction	8
2.2 Information Systems in Educational Settings	9
2.2.1 The Pedagogical Dimension	10
2.2.2 The Technological Dimension.....	11
2.2.3 Outline of the Argument in this Literature Review	12
2.3 Teacher Quality is a Significant Factor in Learning	14
2.4 How to Improve Teacher Quality—Teacher Professional Development	19
2.4.1 AERA President Calls for Situative Research on Teacher PD.....	20
2.4.2 Fishman, Richardson and the KBA Framework	21
2.4.3 Analysis of the Strengths and Weaknesses of the KBA Framework.....	24
2.4.4 Incorporating Research Findings into Professional Practice	25
2.4.5 The Transtheoretical Model of Behavioral Change.....	27
2.4.6 The Fit Between TTM and KBA	33
2.5 Teachers’ Competence at Assessment	34
2.5.1 Types of Practices Exemplifying Formative Assessment.....	39
2.5.1.1 Higher-Order Questions.....	39
2.5.1.2 Providing Useful Feedback.....	42

TABLE OF CONTENTS
(Continued)

Chapter	Page
2.5.1.3 Encouraging Students to Engage in Self-Assessment	43
2.5.2 Attitudes and Beliefs Associated with Formative Assessment Practice	44
2.6 Multiple-Choice Questions.....	46
2.7 Discussion of the Assessment/PD Literature—Implications for System Design	51
2.8 Catalogue of Extant Systems.....	52
2.8.1 Commercial Systems	52
2.8.2 Systems Generated from Basic Research	55
2.8.3 Shareware and Other Tools.....	56
2.8.4 Discussion	57
2.9 Persuasive Technology.....	58
2.10 Technology Shapes Problems	62
2.11 Discussion and Conclusion	64
2.11.1 Developing a system to help teachers get better at assessment	64
2.11.2 The Dissertation	67
3 TWO PILOT STUDIES INVESTIGATING MULTIPLE-CHOICE QUESTIONS ...	68
3.1 Introduction	68
3.2 The QuizViz Study	69
3.2.1 Description of the QuizViz Study.....	69
3.2.2 Results.....	71
3.3 The Second Pilot Study	77
3.3.1 Description of the Study	77

TABLE OF CONTENTS
(Continued)

Chapter	Page
3.3.2 Results.....	78
4 AN INFORMATION SYSTEM MODEL FOR TEACHER CHANGE	82
4.1 Introduction	82
4.2 Elements of the Model	83
4.2.1 The Core—KBA	85
4.2.2 The Organizing Framework—The Transtheoretical Model	88
4.2.2.1 Stages of Change.....	89
4.2.2.2 Processes of Change	90
4.2.2.3 Metrics of Change.....	91
4.2.2.4 An Implication of TTM	92
4.2.3 The Education Domain	93
4.2.4 The Information Systems Domain.....	96
4.2.5 The Outside World—Interoperability with IMS Data Models.....	100
4.3 Combining the Elements	101
5 DESCRIPTION OF QUESGEN	103
5.1 Introduction	103
5.2 Implemented Processes of Change.....	105
5.2.1 Consciousness Raising.....	105
5.2.1.1 Tutorials	106
5.2.1.2 Online Help.....	107
5.2.1.3 Screen Designs.....	108

**TABLE OF CONTENTS
(Continued)**

Chapter	Page
5.2.2 Counterconditioning	111
5.2.3 Self-Reevaluation.....	112
5.2.3.1 Educational Objectives	113
5.2.3.2 Item Analysis Reports.....	114
5.2.4 Environmental Reevaluation.....	114
5.2.4.1 Item Analysis Reports.....	115
5.2.5 Self-Liberation	115
5.2.5.1 Choice of Templates	116
5.2.5.2 Custom Templates	116
5.2.6 Social Liberation.....	117
5.2.6.1 Printing Questions.....	117
5.2.6.2 Export for WebCT	118
5.3 Non-Implemented Processes of Change.....	118
5.3.1 Helping Relationships.....	119
5.3.1.1 Online Forums	119
5.3.1.2 Peer Review	120
5.3.1.3 Question Exchange	120
5.3.2 Stimulus Control.....	121
5.3.3 Contingency Management	121
5.4 Data Model and Scope of Functionality.....	122
5.5 Conclusion.....	125

TABLE OF CONTENTS
(Continued)

Chapter	Page
6 OPERATIONALIZING QUESTION QUALITY	126
6.1 Introduction	126
6.2 What is a “good” question?	127
6.3 Quantitative Measures of Question Quality	128
6.3.1 The Item Discrimination Index (DI)	128
6.3.2 Logistic Models Derived from Item Response Theory (IRT)	131
6.3.3 Discussion of DI, IRT, and QuesGen	136
6.4 Item-Review Panels	142
6.5 Direct Feedback from Students	145
6.6 System Usage Logs	146
6.7 Interviewing the Teachers—Item Review Sessions	146
6.8 User Satisfaction	149
6.9 Other Interpretations of “Quality”	151
6.10 Summary	154
7 QUESGEN EVALUATION	155
7.1 Introduction	155
7.2 Hypotheses	155
7.2.1 New Functionality	155
7.2.2 User Satisfaction	157
7.2.3 Intervening Variables	159
7.2.4 Student-Related Variables	161

**TABLE OF CONTENTS
(Continued)**

Chapter	Page
7.2.5 Exploratory Questions	162
7.3 Conclusion.....	162
8 EXPERIMENTAL DESIGN.....	163
8.1 Introduction	163
8.2 Variables.....	163
8.2.1 Independent Variables	163
8.2.2 Dependent Variables.....	165
8.3 Control.....	165
8.3.1 Students.....	166
8.3.2 Course Difficulty	166
8.3.3 Differences in the User Interface	167
8.3.4 Differences in Instruction and Content.....	168
8.4 Selecting Courses and Participants.....	169
8.5 Experimental Procedure	171
8.5.1 Week 1: Instructors Write Questions.....	171
8.5.2 Week 2: Students Take Quizzes and Surveys.....	173
8.5.3 Week 3: Expert Panel Reviews Questions.....	174
8.5.4 Week 4: Follow-up Interviews with Instructors	176
8.6 Data Analysis	177
9 RESULTS	178
9.1 Introduction	178

TABLE OF CONTENTS
(Continued)

Chapter	Page
9.2 Experiment Overview.....	180
9.3 Limitations of the Study Design.....	182
9.4 Impact of QuesGen Functionality	185
9.4.1 Aligning Objectives with Questions	186
9.4.2 Using Question Templates.....	187
9.4.3 Using the Question Quality Checklist.....	189
9.5 QuesGen and the Discrimination Index (DI)	191
9.6 Behavioral Intention to Use QuesGen.....	193
9.7 Instructor Experience and Course Differences.....	195
9.7.1 Different Amounts of Experience	195
9.7.2 Different Course Subjects	197
9.8 Students' Evaluations of the Questions.....	199
9.8.1 QuesGen's Impact on Perceptions of Difficulty, Clarity and Fairness	200
9.8.2 Impact of Instructor Experience on Difficulty, Clarity, and Fairness..	202
9.8.3 The Impact of Course Content on Difficulty, Clarity, and Fairness	203
9.9 Summary	204
10 DISCUSSION	205
10.1 Improving Question Quality with QuesGen?.....	205
10.2 Why QuesGen Didn't Foster Alignment of Questions to Objectives	206
10.3 What the Item-Review Instrument Says About Question Quality	209
10.4 Interpreting the Discrimination Index	214

TABLE OF CONTENTS
(Continued)

Chapter	Page
10.5 Course and Experience Effects.....	221
10.6 Is QuesGen Effective?.....	226
11 CONCLUSION AND FUTURE WORK.....	233
11.1 The Context of This Dissertation	233
11.2 Contributions	234
11.2.1 New, Web-based Software Tools for Writing and Evaluating MCQs. 234	
11.2.2 A Clearer Picture of the Dimensions of MCQ Quality.....	234
11.2.3 A New Model for Teacher Professional Development Software	235
11.2.4 Data-Gathering Integrated Into System Design.....	235
11.2.5 Instruments for Understanding Question Quality	236
11.3 Future Work	236
11.3.1 QGv2.....	237
11.3.2 Research.....	239
11.4 Conclusion.....	241
APPENDIX A STUDENT CONSENT TO PARTICIPATE IN RESEARCH PROJECT	242
APPENDIX B INSTRUCTOR CONSENT TO PARTICIPATE IN CONFIDENTIAL RESEARCH.....	246
APPENDIX C ITEM REVIEW INSTRUMENT USED BY EXPERT PANEL	250
APPENDIX D QUESGEN: QUESTION ANALYSIS REPORT (EXAMPLE)	253
APPENDIX E FOLLOW-UP INTERVIEW GUIDE.....	255
APPENDIX F STUDENT QUESTIONNAIRE	258

TABLE OF CONTENTS
(Continued)

Chapter	Page
APPENDIX G INTRODUCTORY EMAIL SENT TO PARTICIPATING INSTRUCTORS.....	259
APPENDIX H INTRODUCTORY EMAIL SENT TO STUDENT PARTICIPANTS	262
APPENDIX I UTAUT INSTRUMENT USED IN EVALUATING QUESGEN.....	264
REFERENCES	266

LIST OF TABLES

Table	Page
3.1 The QuizViz Questionnaire	70
3.2 Frequencies of Quiz Questions in Bloom’s Taxonomy Categories (N=40).....	74
3.3 Results from Quiz #1 as Reported by WebCT™.....	74
6.1 Potential Measures of QuesGen's Effectiveness	151
9.1 Synopsis of QuesGen Study Results.....	179
9.2 Breakdown of QuesGen Participating Instructors	180
9.3 Distribution of Instructors by Course and Question Writing Experience.....	184
9.4 Objectives Selected with QuesGen vs. Judge’s Evaluation that Objectives were Addressed by the Question	186
9.5 Results of Chi-squared Tests of Adherence to Objectives Controlling for Instructor Experience and Course Taught.....	187
9.6 Template Selected with QuesGen vs. Judges Evaluation that Questions Addressed Cognitive Skills Higher than Recall.....	188
9.7 ANOVA Results for Impact of Course and Instructor Experience on Behavioral Intention to Use QuesGen	195
9.8 Contribution of Course and Instructor Experience	195
9.9 Relationship Between Instructor Experience and Judges’ Evaluations that Questions Assessed Higher-Order Bloom’s Taxonomy Levels	196
9.10 Relationship Between Course Subject and Judges’ Evaluations that Questions Assessed Higher-Order Bloom’s Taxonomy Levels	197
9.11 Results of Significance Tests for Satisfaction Differences Between GCOM and GKIN Instructors.....	199
9.12 Results of Tests of QuesGen’s Impact on Student Perceptions.....	200
9.13 Results of Tests of Impact of Instructor Experience on Student Perceptions.....	202
9.14 Results of Tests of Impact of Course Content on Student Perceptions	204

**LIST OF TABLES
(Continued)**

Table	Page
10.1 Factor Loadings for Varimax Rotation of Item-Review Instrument Items.....	211
10.2 Results of Tests for Relationships Between Factors and Independent Variables	213
10.3 Correlation Matrix of DI with Factors from the Item-Review Analysis	216
10.4 Distribution of Instructors by Course and Question Writing Experience.....	221
10.5 Correlation Matrix for Items on the Item-review Instrument	232

LIST OF FIGURES

Figure	Page
1.1 Sample of a statistical report produced by WebCT	1
2.1 The Knowledge, Beliefs, and Attitudes (KBA) Framework from Kubitskey and Fishman (2005)	21
2.2 The Role of Assessment in Effective Instruction	38
2.3 Steps in Professional Item Development (Haladyna, 2001, p4)	48
2.4 Revised Bloom's Taxonomy of Educational Objectives (Anderson and Krathwohl, 2001). Shaded region indicates areas Haladyna (2001) considers appropriate for assessment with MCQs.	49
2.5 Guidelines for MCQs (Haladyna, et al., 2002)	50
3.1 Bad Question Resulting from Assessment of Multiple Concepts.....	75
3.2 A rewrite of the previous question to correct ambiguity problems	75
3.3 Bad Questions Resulting from Poor Exemplification of Concepts	76
4.1 Elements of the system model	83
4.2 A System Design Team for a PD System	84
4.3 The Expanded System Model	102
5.1 A Screenshot of a QuesGen Video Tutorial	106
5.2 A Screenshot of an Online Help Popup Window	107
5.3 Question Quality Checklist	108
5.4 Screen Design of the Question Entry Form	109
5.5 The Question Entry Form with Enlarged Fonts	111
5.6 Encouraging Use of Objectives	112
5.7 Objectives Offer a Chance to Reflect	113
5.8 Using Templates for Reasoning Questions	116
5.9 Export Questions for WebCT	118

**LIST OF FIGURES
(Continued)**

Figure	Page
5.10 QuesGen System Boundary. Adapted from IMS QTI Overview, p5.	122
5.11 QuesGen IMS-compatible Data Model	124
6.1 Sample Item Characteristic Curves (ICC) (image from http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm).....	132
6.2 Sample ICC Curves for the 3-parameter Logistic Model (image from http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm).....	134
8.1 The QuesGen Interface	167
8.2 The "Standard" Interface.....	168
9.1 Question Exhibiting Formatting Problems Instructor’s data was removed from the study—all answer options are given as the first choice	181
9.2 Distribution of DI Scores for Different System User Groups.....	191
9.3 Satisfaction Scores for All Participants (N=17)	198
10.1 Educational Objectives from the Two Courses	223
10.2 Hypothetical Graph of Improvement in Instructor MCQ Quality Over Time With and Without Training	225

CHAPTER 1

INTRODUCTION

1.1 Motivation

In the fall of 2002 ago a subset of students and faculty in the department began studying the field of information visualization. Attendees of the seminar read a collection of articles on the subject compiled by the foremost experts in the area. After reading these articles this new understanding was to be applied to an information problem, and so began the search for an information system with an interface that could benefit from augmentation with an alternative way to visualize the information displayed by that system. At the time, one of the seminar members was using WebCT, a popular commercial learning management system (LMS), to teach a course, and had been experimenting with the quiz tool that is built into the system. One of the features provided by WebCT's quiz tool is automatically generated statistical reports on the performance of multiple-choice questions (MCQs) used on a quiz (see Figure 1.1). These statistical representations became the focus of study.

Title	N	% Correct Of:			Discrimination	Score	
		Whole Group	Upper 25%	Lower 25%		Mean	SD
Groupware- General	22	50	80	0	0.63	50.0%	51.2
CSCW - General	22	100	100	100	0.00	100.0%	0.0
Workflow Computing - General	22	81	90	66	0.26	81.8%	39.5
Montage Video	22	22	40	16	0.54	22.7%	42.9
Scientific Collaboration Study	22	9	20	0	0.64	9.1%	29.4
Information Visualization - General	22	95	100	83	0.43	95.5%	21.3
Pre-attentive processing-Definition	22	36	60	16	0.51	36.4%	49.2
Pre-Attentive Features	22	90	100	83	0.26	90.9%	29.4
Visual Pattern Techniques	22	68	100	0	0.68	68.2%	47.7
3D-Techniques	22	90	100	83	0.36	90.9%	29.4
Overall Mean:						64.5%	

Figure 1.1—Sample of a statistical report produced by WebCT

An early thought was to take this representation and convert it into a format that would be more useful to teachers. For example, not having training in psychometrics, teachers are not familiar with the “discrimination” statistic that was reported in the table, nor do they have a firm grasp of how to interpret it in the context of quiz questions. The question was asked whether there might be other better ways to describe the performance of the quiz questions to the instructors who were using them. This led to an exploratory study (described more fully in Chapter 3), in which the students taking a quiz gave their feedback on the difficulty, clarity, and fairness of the quiz questions. An unexpected finding was that a surprising number of the questions on the quiz were of poor quality. The results of this study were written up and submitted to a conference, and the focus of research was no longer on helping teachers visualize the results of a quiz, but on how to help teachers get better at writing questions in the first place.

1.2 Research Questions

A web-based tool, called QuesGen, was built to help teachers write better multiple-choice questions. Grounded in the research to be reviewed in the next chapter, QuesGen was motivated by two key observations:

1. The quality of assessments used to gauge student learning contributes significantly to what and how much students learn, and
2. Teachers are not very competent at creating high-quality assessments.

QuesGen was built to augment or replace the types of MCQ-writing tools currently available to teachers. With current tools, teachers can enter questions, randomize the order of the answer choices, include images and mathematical formulas,

and have flexible options for online delivery and scoring. A major assumption of these tools is that *teachers already know how to write good multiple-choice questions*. The literature indicates otherwise, however, and as such, QuesGen's contribution to this area is to be the first tool that provides concrete guidance on how to write good questions. This guidance is realized via three new sets of functionality.

1. A mechanism for entering and explicitly aligning questions with educational objectives,
2. Semantic question templates designed to jumpstart teachers' creative process, encourage greater question-type diversity, and focus questions on higher-order thinking, and
3. A question-quality checklist similar to those employed by item-review panels in the professional test creation industry

Having this new set of functionality allows a number of research questions to be asked. These research questions can be broken up into two groups, one set which looks at the impact of the new functionality on question quality, and a second set which examines the user interface and its role in how teachers experience the new functionality.

The first set of questions breaks down the larger, more general question of whether or not using QuesGen improves the quality of MCQs. Research questions are posed for each of the three new types of functionality:

RQ1: Does the explicit association of an educational objective with a multiple-choice question increase the chance that the question will assess mastery of one of the stated objectives of the given unit of instruction?

RQ2a: Does the inclusion of semantic question templates result in a greater diversity of MCQ types that a teacher will write?

RQ2b: Does the inclusion of semantic question templates result in a greater number of questions targeting cognitive skills greater than recall, i.e. understanding, application, analysis?

RQ3: Does the inclusion of a question quality checklist lead teachers to write questions with fewer technical flaws than without the checklist?

Before going on to introduce the second set of research questions, the rationale for the above questions will be briefly discussed. The literature review in Chapter 2 will add more depth and context to this discussion.

The motivation for RQ1 is best captured with the well-known saying, “If you aim at nothing, you will surely hit it every time.” It rests on the observation that when objectives are not explicitly stated, MCQs tend to focus on irrelevant or trivial knowledge. Teachers may assume that they have a salient objective in mind while writing a question, but when asked to work backward from a question to determine what objective it assesses, teachers are often frustrated to find that their questions assess items that are not that important. A simple way to address this issue is to provide teachers the opportunity to incorporate objectives into the question. QuesGen would be the first tool to do this.

Questions RQ2a and RQ2b deal with the issue of cognitive complexity. A major goal of education is not just to transfer knowledge, i.e. things to be remembered, but also to teach students how to think. Questions that ask students to understand concepts, apply procedures, or analyze information are more cognitively complex than questions that only ask students to remember facts or information. It is desirable to have a set of questions

that addresses a range of different cognitive levels, as well as targeting behaviors that are more cognitively complex than just recall. QuesGen does not include any templates which focus only upon recall.

Often the problems with MCQs come from spelling or typographical errors, the inclusion of more than one, or no correct answers, vocabulary that is too difficult for the students, or other documented and avoidable technical flaws. In the psychometric industry, item-review panels examine questions to make sure that they don't contain such flaws. Many teachers, however, are unaware of the flaws and do not have the training to find them. Since most of the flaws are relatively straightforward, the inclusion of a checklist in QuesGen gives teachers an opportunity to proofread their questions in the same way that professional item writers do.

Taken individually or as a set, it was believed that the functionality described above would be effective at improving the quality of MCQs for the reasons discussed by the research questions. However, understanding the importance of objectives, specifying objectives, learning how to make use of the templates and the checklist are all activities that require time and effort. Incorporating the functionality above into an interface that convinces instructors to take the time necessary to improve their skill at writing MCQs is a non-trivial task. As such, the second set of research questions deals with the MCQ-writing experience and seeks generally to gain insight into the role of the interface in mediating the effectiveness of the new functionality.

RQ4: What will teachers relative level of satisfaction be with the QuesGen tool compared to a tool without the new functionality?

RQ5: Relative to a tool without QuesGen functionality, are teachers likely to say that they will use QuesGen for writing multiple-choice questions?

RQ6: Does the inclusion of semantic question templates decrease the time—real or perceived— it takes for a teacher to write a new question?

Finally, there are some research questions that are of interest either because they deal with an intervening variable affecting QuesGen’s performance, or because they are exploratory.

RQ7: How will interaction with QuesGen impact teachers’ attitudes toward using MCQs for assessment?

RQ8: How will the impact of QuesGen differ across different subject areas?

RQ9: What is the role of teachers’ experience in the resulting quality of questions?

The remainder of this dissertation explores these research questions, gathers evidence, analyzes the resulting data and draws conclusions about QuesGen’s effectiveness and the path forward.

1.3 Organization of the Dissertation

This thesis is organized into eleven chapters. Chapter 1, which you’re reading now, introduces the original motivation for this research, the primary research question, and the outline for the thesis. Chapter 2 explores and analyzes the literature relevant to answering the research question. Chapter 3 describes two pilot studies that informed the design both of the software tool that was built, and of the evaluation strategy that was used to test it. Chapter 4 synthesizes the results of the first pilot study with the various

ideas found in the literature review to develop a generalized information systems model whose goal is to guide the development of systems to foster the development of teachers' professional skills. Then in Chapter 5 a detailed description is given for QuesGen, which was built following the information systems model. Chapter 1 is an in depth discussion of how to operationalize the concept of question quality. Chapter 7 presents and justifies the hypotheses that were researched using the experimental design described in Chapter 8. Chapter 9 presents the results of the experimental study. Chapter 10 is a discussion of those results, and the final chapter, Chapter 11, draws conclusions and lays out plans for future work.

CHAPTER 2

THE STATE OF THE ART

2.1 Introduction

This literature review explores the role of information systems in improving teachers' competence at educational assessment. Recent research in education indicates that the quality of teaching has a very significant impact on learning when compared to other factors in learning environments. Improving teaching is therefore likely to be a productive means of achieving learning gains in students. This research targets teachers' competence at assessment as a key skill area that could be improved in teachers. The goal of this chapter is to use existing published research to both help define the requirements of an information system that might help teachers to become better at assessment and to develop an experiment that will evaluate the effectiveness of this information systems aid. In effect, this chapter defines the state of the art in this type of system. Several relevant areas of literature occupy this information space. First is research on the successful strategies for achieving teacher change that comes from research on teacher professional development. From this literature it is clear that the system that will effect teacher change needs to be more than a tool—such a system must actually persuade teachers to undertake the effort necessary to improve. Thus the literature on persuasive information systems will be examined. Furthermore, since a review of effective assessment practices indicates a specific set of practices, the literature on information systems as defining processes that structure problems will also be examined. This work would also not be complete without a review of extant assessment

support systems. This chapter will catalogue, classify, and discuss the strengths and weaknesses of the systems currently employed to support teachers' assessment practices and conclude by describing applicable research methodologies that measure the effectiveness of teacher support systems. The chapter will conclude with a discussion showing that research can be done to provide information on how to make these systems more effective.

2.2 Information Systems in Educational Settings

If the broad goal of information systems is to improve both the efficiency and effectiveness of human endeavor, then the goal of information systems in educational settings is to improve both the efficiency and effectiveness of learning. A principal contribution to the theory of information systems in educational settings comes from Leidner and Jarvenpaa, who reported in 1995 that information technology within management education was not fulfilling its role, saying:

Our analysis suggests that initial attempts to bring information technology to management education follow a classic story of automating rather than transforming. IT is primarily used to automate the information delivery function in classrooms. In the absence of fundamental changes to the teaching and learning process, such classrooms may do little but speed up ineffective processes and methods of teaching (Leidner and Jarvenpaa, 1995, p265).

Their analysis goes on to categorize the types of technology being used with various learning and pedagogical styles, and to discuss how each category of technology interacts with each of the learning styles. Their premise is that the effectiveness of the technology will be a function of the fit between the technology and the type of learning called for in a given situation. This premise strongly parallels that of the task-technology

fit (TTF) model proposed by Goodhue and Thompson, also in 1995 (Goodhue and Thompson, 1995). The two main dimensions of Leidner and Jarvenpaa's model are the pedagogical dimension, and the technology type dimension.

2.2.1 The Pedagogical Dimension

The pedagogical dimension of the taxonomy is divided into five learning models: objectivism, constructivism, collaborativism, cognitive information processing, and socioculturalism. Of these five, only the first four are found in general practice in classrooms. Socioculturalism (O'Loughlin, 1992) is a political reaction to what is seen to be the enforcement of cultural hegemony by the other learning models, and does not have specific technology nor a well-developed pedagogy associated with it. As such, this view does not figure heavily into Leidner and Jarvenpaa's taxonomy. Figuring more heavily into the model are the other four models. The objectivist model assumes that the objects of learning are fixed and reside within the mind of the instructor. Learning means a transfer of this knowledge from instructor to student, and learning is assessed primarily by means of checking students' ability to recall this knowledge upon command. The constructivist model posits that students learn by participating in guided experiences and then constructing knowledge on their own by "connecting the dots" between these various experiences. Assessment of learning in this model involves having students interpret the meaning of knowledge with respect to the context of their experiences. Collaborativism advances the notion that understanding about the world arises through interaction with others which creates shared conceptions of reality. Activities are heavily focused on group work, and assessment may also rest largely on the evaluation of the products of groups of students, rather than individuals. Finally, the cognitive information

processing model states that learning is the process of transferring new knowledge into long-term memory. Instructors must closely monitor students' level of understanding and adjust the stimuli in their environment to bring about desired learning gains. Each of these learning models matches up roughly with one or more of four styles of information technology for learning.

2.2.2 The Technological Dimension

The four styles of learning technology are automation, informing-up, informing-down, and transformation. Automation takes traditional forms of learning materials and uses IT to partially or wholly automate their delivery. In automation, PowerPoint replaces the blackboard, and online exercises replace those traditionally performed from a textbook or in a notebook. The use of networks to permit distance learning is also classified here. Automation is roughly associated with objectivist pedagogy. Informing-up refers to systems that provide the instructor with more, and more detailed information about the current state of student learning with the expectation that the instructor can then use this information to tailor and revise instruction on-the-fly to meet student needs. Electronic keypads that allow in-class voting or question-answering, as well as e-mail are examples of technologies that fit this category. The responsiveness of these technologies is seen to fit best with the cognitive information processing model of learning. Informing-down means providing learners with rich, interactive information that can allow them to explore more fully a given topic area and take more control over their own learning process. Examples of IT in this category are simulations, hypermedia, learning networks, groupware, and other forms of computer-mediated communication (CMC). Technologies which are focused more towards individuals such as simulations and hypermedia, are

considered to support the constructivist model, whereas technologies that foster collaboration, such as groupware, are classified as supporting the collaborativist model. Finally, the transforming role of IT in learning would dramatically alter the roles and hierarchy of participants in the learning process. Transforming IT appears to be similar to technology for informing-down, however the key difference is that learning groups in transforming settings would last much longer and make much higher use of the asynchronous aspects of the technology. Influence in this structure would stem from expertise and would not rest necessarily with a single person designated as the “instructor.”

In noting the strengths and weaknesses of Leidner and Jarvenpaa’s framework, it is important to note that this work was done prior to the Internet boom of the mid to late 1990’s, and as such, many of the technologies they discuss were still in their infancy. Many others, such as instant messaging, “podcasts,” and advanced wireless and cell-phone-based technologies, did not exist yet. A strength of the model is that not only is it possible to place these new technologies into their framework, but also the consideration of these new technologies with respect to the framework suggests possibilities and uses for them in educational settings not thought of before. On the other hand, Leidner and Jarvenpaa’s framework doesn’t go as far as describing what “fundamental changes” need to occur to improve learning, or specifically how technology might bring these changes about.

2.2.3 Outline of the Argument in this Literature Review

This State of the Art literature review will explore the question of what is the most effective way for technology to improve learning, and it will begin to flesh out a design

for a novel information system designed for this purpose. In pursuing this question this literature review will address one of the six avenues for future research described by Leidner and Jarvenpaa, namely their indication that:

Research is needed on understanding the roles of instructors and students as well as the appropriate learning assessment strategies in virtual learning spaces (Leidner and Jarvenpaa, 1995, p287)

The role of the teacher in fostering learning will be reviewed, calling attention to recent education research highlighting the importance of teachers. This focus leads into a discussion of the research on how to help teachers improve, and hence, a review of the literature in teacher professional development. At this point in the narrative, a detour is taken to examine the transtheoretical model, a promising line of research from the health behaviors literature which may provide guidance into the best way to design intervention programs to increase teacher competence. Having gained an understanding of how to help teachers improve, this review will next turn to the issue of in which area teachers most need improvement. This discussion will focus on the area of teachers' competence at classroom assessment techniques, and even more specifically focus on the topic of teachers' development and use of multiple-choice questions—an area in which it will be argued that teachers have strong need for improvement, and which, if attained, will significantly improve student learning. At this point, the extant information systems designed to support teachers' development and use of multiple-choice questions will be examined and shown to be weak in precisely the way that Leidner and Jarvenpaa describe, i.e. they merely automate and do not transform teacher practice. The last sections of this review will turn to the challenge of designing an information system that will aid and encourage teachers to develop and use better MCQs. Since the system will attempt to persuade users to adopt new behaviors, the literature on persuasive technologies will be

reviewed. Also since the way that teachers design MCQs may have an impact on the way that they structure their lessons, the literature on structuration theory—how information systems structure problems—will be reviewed. Having covered this ground, this literature review will prepare the way for the design and testing of an actual system to help teachers write better MCQs. This system and its evaluation are described in the chapters that follow.

2.3 Teacher Quality is a Significant Factor in Learning

Does the teacher matter? While few would dispute that students learn more with a teacher than without one, with all of the other factors that affect learning—the student’s academic ability, home environment, socio-economic status, the curriculum and facilities provided by the school system—it’s very difficult to know how much individual teachers contribute to students’ learning (Sanders and Horn, 1994). A number of factors over the past several decades have made this an ever more important question: namely increased understanding of the role of tutors in learning (Bloom, 1984; Chi, et al., 2004; Chi, et al., 2001; VanLehn, et al., 2003), the sophistication of standardized testing (psychometrics) and statistical methods for isolating teacher contributions to learning (Campbell, et al., 2004), all of which have encouraged policy makers to enact increasingly ambitious school accountability statutes. The argument is made in this section that the quality of teaching students receive is one of the most significant contributors to student learning.

Bloom’s “2-sigma problem” paper (Bloom, 1984) describes data showing that with one-on-one human tutoring, students consistently perform more than two standard deviations better on learning tasks than in group instruction, as in a typical classroom

setting. In the study, students of similar academic ability were randomly assigned to a teaching condition, either traditional classroom instruction or one-on-one tutoring, and instructed over a three-week period. Both sets of students received equal amounts of instruction time. The experiment was repeated with different subject matter and in different grade levels all with comparable results. To summarize, “about 90% of the tutored students...attained the level of summative achievement reached by only the highest 20% of the students under conventional conditions.... Typically, the aptitude-achievement correlations changed from +.60 under conventional...to +.25 under tutoring,” (Bloom, 1984) and tutored students spent greater than 90% time on task compared to only about 65% for conventionally taught students. To use Bloom’s words, the problem raised by this research is:

Can researchers and teachers devise teaching-learning conditions that will enable the majority of students under group instruction to attain levels of achievement that can at present be reached only under good tutoring conditions? (p4-5)

Bloom’s study also explored a third experimental condition in which conventional classroom instruction was replaced with a style of instruction called mastery learning. In mastery learning the instruction is essentially the same as in conventional classrooms, however the main difference is that in conventional settings tests are used only to mark students’ progress, whereas in mastery learning tests are used to provide feedback and guidance to students who then practice the materials and re-test until mastery is achieved. Students in the mastery learning condition showed learning gains on the order of one standard deviation over conventionally taught students, or about half of the gains showed by tutored students. In Bloom’s estimation, the amount of effort required to get teachers to adopt the practices of mastery learning was relatively small, particularly in relation to

the learning gains to be realized from it. In other words, the teachers matter a lot, and one of the best ways to improve learning is to improve the quality of the teachers.

In contrast to Bloom's study, which provides evidence of teachers' importance to student learning on a relatively small scale, Sanders' Tennessee Value-Added Assessment System (TVAAS) provides evidence on a large scale. In the early 1980's William Sanders at the University of Tennessee, Knoxville¹ took on the task of finding a way to use state standardized test data from primary and secondary schools in the assessment of school and teacher quality (Hill, 2000; Sanders and Horn, 1994). The testing regime supporting TVAAS was described by an external validity review panel as follows:

[S]eparate grade-level test booklets are supplied by CTB [California Test Bureau/ McGraw Hill] and administered by TCAP [Tennessee Comprehensive Assessment Program] to all students in grades 2-8 of Tennessee public schools during the last weeks of the school year. The test booklets contain sections devoted to tests in five subject-matters: reading, language, math, science, and social studies. The tests in the first three subject areas contain two types of items: so-called "norm-referenced" items and "criterion-referenced" items. The science and social studies tests contain only norm-referenced items. TVAAS makes use of test scores on the norm-referenced items only. These scores are expressed on a special scale, constructed by CTB, that ranges from 0 to 999 and applies across all grade levels.

Because the successive grade-level tests are reported on a common scale, the scores can be used to measure a student's growth in achievement from one school grade to another. The availability of this type of scale for reporting test performance is essential to TVAAS, which is based on the measurement of annual gains ("value-added") rather than on the test scores themselves. (Bock, et al., 1996, p1)

The scores of each student are tracked across each grade level, and follow the student from school to school throughout the state. The review panel indicates the

¹ William Sanders is now employed by the SAS Institute.

uniqueness of the data available for doing TVAAS-style comparisons in their conclusions as follows:

The educational data collection and management system implemented for TVAAS, in combination with the Tennessee Comprehensive Assessment Program annual achievement testing in grades 2-8, is virtually unique among the states in its ability to keep a continuing record of students' achievement test scores as they move from grade to grade or school to school in each county of the state.(Bock, et al., 1996)

Saunders (1999) provides a history of the use of the term 'value-added' in educational contexts. The term originated in the field of economics and is associated with the transformation of inputs via a process that leads to outputs of higher value. This is relatively straightforward in manufacturing where, for example, a clothing manufacturer starts with raw cloth and converts it into blue jeans. The value added in this case would be the difference between the monetary value of the blue jeans and the cost to the manufacturer of the cloth used to make the blue jeans. In educational terms, the idea is that students enter a classroom with a certain level of knowledge and leave with a higher level and the difference between the two is the "value added" by the educational system. Beginning in the mid to late 1970's, value-added methodologies were largely developed in the 1980's and came into prominence in the early 1990's when policy-makers in the US and UK began to use the results to evaluate schools and school systems for funding purposes (Saunders, 1999).

The TVAAS has received mixed reviews. Professor Sanders worked closely with the Tennessee legislature in 1990 to draft a law which establishes the TVAAS methodology as what will be used in the state to evaluate school performance for resource allocation purposes (Hill, 2000). This led to several state sponsored reviews of the validity of the TVAAS methodology (Bock, et al., 1996; Fisher, 1996; Stroup, 1995),

and has prompted others to write independent reviews (Kupermintz, 2002). In reading these reviews, there is a great deal of concern for the public policy and personnel management issues that they bring up, and these questions are far from being answered (Darling-Hammond and Youngs, 2002). However controversial the application of the results, the reviewers agree that the value-added methodology does indicate that the quality of teachers is important in determining the learning outcomes of students, and the TVAAS literature has been accepted into a growing body of literature which cites the importance of teachers in learning (Darling-Hammond and Youngs, 2002; McCaffrey, et al., 2003; Rowe, 2003).

While Bloom's 2-sigma study and the TVAAS work represent two ends of the spectrum (micro and macro) with respect to the work on quantifying teacher effectiveness, other studies support the focus on teacher improvement as a means to increasing learning. Campbell et al. (2003) tender a definition of teacher effectiveness that while "not perfect" at least serves as a "working definition" to fill the void left by implicit definitions in others' work. They propose that teacher effectiveness is:

the power to realize socially valued objectives agreed for teachers' work, especially, but not exclusively, to work concerned with enabling pupils to learn (p354).

They state that their definition assumes some appropriate means for measuring whether or not the agreed upon objectives have been attained, and they go on to propose differential measures for teacher effectiveness, espousing the position that effectiveness is largely contextual and that broadly generalized models of teacher effectiveness may not be appropriate. All of this discussion rests upon their survey of recent research indicating that teacher effectiveness is a critical factor in student learning.

Campbell et al.'s study along with another by Ellet and Teddlie (2003) both give roughly chronological accounts of research into teacher effectiveness, and agree on the characteristics of modern conceptions of teacher effectiveness. Current conceptions build upon the role of teachers' expectations for students, and upon the depth of understanding and pedagogical knowledge of the teacher. The importance of teachers' expectations is demonstrated, for example, in the work of Brophy (1983). The understanding of teachers' pedagogical knowledge is described in the work of Shulman (1986). The net effect of these studies is that current practices for measuring teacher effectiveness are increasingly learner-focused. Models of effective classrooms revolve around learning activities where students partner in the construction of new knowledge. To tie this work back into the Leidner and Jarvenpaa framework from the introduction, recent work on teacher effectiveness provides increasing support for the constructivist and collaborativist approaches to learning.

2.4 How to Improve Teacher Quality—Teacher Professional Development

If it is accepted that improving teachers is a logical path for improving student learning, the next question becomes: how best can teachers be improved? Answering this question involves answering two sub-questions: what aspect of teachers needs improvement, and what should be the means of doing so? The second question is addressed in this section, and the prior question in the next.

Professional development (PD) in any field is primarily concerned with changing behavior. More specifically in a teacher professional development context, PD is concerned with getting teachers to replace ineffective instructional methods with

effective ones, or to acquire new behaviors that are designed ultimately to increase student learning. Effecting behavioral change on a large scale, i.e. in all teachers, is a task with many complicating factors.

2.4.1 AERA President Calls for Situative Research on Teacher PD

Hilda Borko (2004), former president of the American Educational Research Association, used her presidential address at the AERA's annual meeting to address the topic of research into teacher professional development. She echoed what others have said, that teacher learning is key to teacher development (Fishman, et al., 2000; National Council of Teachers of Mathematics, 2000; National Research Council, 1996; Richardson, 2003b; Sherin, 2002; Shulman, 1986). A key question that she raises is how does one go about doing the research that will best reveal how to help teachers learn to do their jobs better?

Borko advocates a situative approach to research, which “allows for multiple conceptual perspectives and multiple units of analysis,”(Borko, 2004, p4). By multiple perspectives, the situative approach seeks to integrate and leverage the insights of individual behavior gleaned from psychological research, with understanding of group behavior from more macro level sociocultural research. Such research results in professional development paradigms which address issues both at the individual teacher and student level, and at the class, school, and school system level, and which take into account various factors such as teacher and student cognition, motivational and political pressures, organizational management, and long-term sustainability. Such a comprehensive approach has been put forth by Fishman, et al. in answer to Borko's call.

2.4.2 Fishman, Richardson and the KBA Framework

Fishman et al. (2003) maintain that the lack of empirical evidence regarding what teachers actually learn as a result of professional development interventions makes it difficult or impossible to make intelligent choices about the design of those interventions. They propose a framework that is used to guide systematic implementation of professional development programs and also collect empirical data on the quality of those programs. Their framework was first introduced in Fishman, et al. (2000) and has been further refined in Kubitskey and Fishman (2005) (see Figure 2.1).

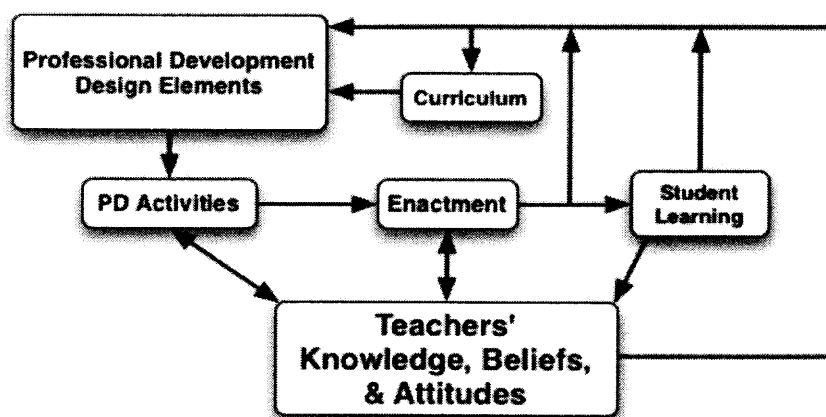


Figure 2.1—The Knowledge, Beliefs, and Attitudes (KBA) Framework from Kubitskey and Fishman (2005)

The goal of the model is to produce high-quality professional development programs which they describe as follows:

[High quality professional development is] well planned over the long term using evaluations of past PD opportunities to inform future PD. The PD is structured around a specific need of participants, proximal to teacher practice and supplying usable information to teachers. The participants should either already be in a community or should have common ground for forming a community, such as teaching the same unit, teaching in the same school, etc. The PD should take place over an extended period of time, to promote the creation of a community of practice whose participants have common goals addressed by the PD. Finally, the PD activities themselves must offer opportunities for teachers

to engage in inquiry of either content or pedagogy.(Kubitskey and Fishman, 2005, p2)

The core assertion of the model builds upon Richardson's (1996) finding that classroom practices are strongly correlated with teachers' knowledge, beliefs and attitudes (KBA); hence the goal of professional development should be to modify these three teacher attributes. Professional development is described as a cycle in which teachers enact changes in their classroom practices as a result of professional development activities. Their enactment impacts KBA, which in turn impacts the enactment. The results of enactment also impact student learning which also feeds back into KBA. All three of these elements—enactment, KBA, and student learning—impact the curriculum and the elements chosen for the next round of professional development.

The KBA framework addresses a key problem in the field of teacher professional development: it is difficult to build a coherent critical mass of empirical evidence on effective professional development when the range and types of behaviors espoused by PD programs varies so widely. Phrased another way, if there were a relatively small number of clearly identifiable, well defined practices that constituted "best practices" for teachers, measuring them would be a much more straightforward activity and progress in PD research would be much faster. Such a set of teaching practices does not exist. There are many ways in which teachers can be effective and there is not a high degree of consistency across disciplines as to what constitutes the most effective way to teach various subjects (Richardson, 2003b). The critical contribution of the KBA framework is the realization that despite the plurality of effective practices that manifest themselves, there is a relatively stable and small set of knowledge, beliefs, and attitudes that correlate

with effective teaching. After an extended field study of successful PD interventions with secondary school teachers, Black et al. (Black, et al., 2003) summarized that:

Specifically what changed for teachers was

- *their views about learning,*
- *their professional priorities,*
- *their expectations of their students, and*
- *their feelings about control in their classrooms.*

(p91, bullets added for emphasis)

It may be that more than anything else, effective teaching is an outward expression of a core belief system about students, how they learn, and the purpose of a school education. An understanding of these beliefs gives one a handle on how to measure change toward adoption of effective teaching practices—rather than outward behaviors, perhaps the focus of PD research should be on the knowledge, beliefs, attitudes of teachers, and then the ingenuity and effort of teachers in classrooms can be relied upon to bring about learning gains in classrooms.

The KBA methodology is demonstrated in the presentation of case studies of five teachers who were among twenty-eight that participated in a PD workshop designed to teach teachers the value of concept maps as a pedagogical tool, and also how to use them in their classrooms (Kubitskey and Fishman, 2005). Twenty-three of the twenty-eight teachers were interviewed prior to and following the PD workshop and an opportunity to teach a content unit based upon the methodology taught in the workshop. In addition, the teachers were observed in the process of teaching and their students took pre- and post-unit tests on the content material so that learning gains could be tracked. In short, the study found that teachers' knowledge and beliefs with respect to concept mapping as a teaching strategy changed and that these changes made a significant and predictable impact on student learning.

Knowledge was observed corresponding to Shulman's (1986) concept of pedagogical content knowledge. A number of the teachers indicated that they knew about and understood the use of "concept maps" for teaching prior to the PD workshop which was designed to improve teachers' understanding of this pedagogical tool. Pre-workshop interviews indicated that more often than not, teachers' understanding of "concept maps" was not aligned with the version that was to be taught in the workshop. Rather than tell them their understanding was wrong, the workshop introduced a "new way" to use concept maps. Post-workshop interviews and observations of classroom teaching indicated that the workshop changed the teachers' understanding and had an impact on their classroom practices.

In addition, the workshop had an impact on the teachers' beliefs about the effectiveness of concept maps as a pedagogical tool, and about their beliefs that they personally could use them effectively in their instruction. In general, participation in the workshop convinced teachers not only that concept maps are an effective tool, but also that they had the ability to implement them effectively; however, in one case, the teacher's confidence in her ability to use the new tool decreased following her experience with the students. This teacher had strong negative past experiences with elements of the concept map pedagogy, and the researchers concluded that the PD workshop experience was not powerful enough to overcome these negative experiences.

2.4.3 Analysis of the Strengths and Weaknesses of the KBA Framework

To summarize, the KBA framework has a strong empirical and theoretical footing, but may suffer long-term from a problems with scalability. The study described above found strong support for the characteristics of high quality professional development put forth

in the KBA model. To reiterate, high quality PD is long-term, fosters a community of practice, is proximal to teachers' experience and needs, and offers opportunities for teachers to implement and develop their skills immediately in the classroom. Positive change in teachers is associated with the acquisition of pedagogical content knowledge and confidence in the ability to use that knowledge in their classrooms. Their knowledge and beliefs are reinforced or weakened depending on the success of their students in actual learning tasks. The primary problem with the approach described here is that the methods of study are primarily ethnographic meaning that a great deal of effort is required to follow a relatively small number of teachers in necessarily limited subject areas. A possible solution to this problem presented itself when investigating the literature on how to incorporate the findings of research into practice.

2.4.4 Incorporating Research Findings into Professional Practice

Walter et al. (2003) developed a taxonomy of interventions used to increase the impact of scientific research on professional practice. This taxonomy was motivated by the observation that much research on professional practice will not be implemented without explicit strategies for doing so. The taxonomy classifies interventions along two dimensions: type and mechanism. Although the taxonomy identifies thirty-two intervention types, broadly speaking these fall into two categories—presentational (e.g. seminars, in-service training), and collaborative (e.g. partnerships, consortia, practitioner-research). The seven categories of mechanisms for increasing research impact on practice were: dissemination, education, social influence, collaboration, incentives, reinforcement, and facilitation. Included as an eighth “category” of mechanism was “multifaceted initiatives” that employ more than one of the other seven mechanisms. The

paper cited research exemplifying and supporting/decrying the use of various type-mechanism combinations in various contexts. One citation exemplifying a multi-faceted initiative was particularly salient for teacher professional development.

Walter cites Smith (2000) who describes the challenges faced when trying to get physicians to adopt new professional behaviors. Effecting behavioral change in doctors is particularly difficult for several reasons. First, doctors undergo intense and highly proscriptive behavioral modification during medical school, internship, and residency, which establishes a strong pattern from which they are reluctant to deviate. Second, practicing doctors are inundated with invitations to professional development seminars and other opportunities to learn and adopt “cutting edge” technologies/medications/therapies/treatments, frequently sponsored by the purveyors of those technologies/medications/therapies/treatments. Combined with the third factor, that doctors tend to be exceedingly busy, a common defense mechanism is to ignore many or all of these opportunities, and to stick to the practices learned in medical school. Given the frequent life or death nature of the decisions that doctors must make, they must also be extremely cautious and give considerable study to any proposed change in their accustomed practice. Given these pressures, Smith reviews the research on efforts to change doctors’ behaviors, concluding that such efforts are complex and should be more theory-driven.

It is not difficult to draw parallels between Smith’s characterization of doctors and the situation of teachers. By the time they begin teaching they have more than sixteen years of implicit (via their experience as students) and explicit (via their formal teacher training) instruction on how teaching is supposed to be practiced. Teachers are very busy,

and are under a great deal of pressure from parents, their schools, school systems, and state and national statutes—e.g. the No Child Left Behind (NCLB) Act of 2001—to discover and adopt practices which will bring about desired learning outcomes for students (Kyriacou, 2001; Yamagata-Lynch, 2003). Given these similarities, there is an opportunity to apply the ideas in one line of research cited by Smith, namely the transtheoretical model of behavioral change (Prochaska and Velicer, 1997).

2.4.5 The Transtheoretical Model of Behavioral Change

The transtheoretical model (TTM) was developed to address health behaviors, such as smoking, and claims to be the only model of behavioral change that incorporates a temporal dimension. Whereas other models identify behavioral change as a more or less instantaneous event, i.e. the moment someone quits smoking or starts exercising, only TTM identifies change as a process that happens over time. It identifies six stages of change:

- **Precontemplation**

This is the period of time before a person is even considering change. During this phase, a person has no plans to change a “problem” behavior.

- **Contemplation**

Operationalized as the six-month period prior to making a major behavioral change, during the contemplation phase, a person is engaged in weighing the pros and cons of changing. This may or may not involve actively seeking information, or talking with others about these pros and cons. The six-month timeline is measured using a self-report questionnaire.

- **Preparation**

Operationalized as the month prior to making a behavioral change (self-report), a person may engage in activities such as making detailed plans for change, enlisting the support of others, modifying one's environment. It is characterized by a heightened awareness of the pros of changing and the cons for not changing one's behavior.

- **Action**

Experts in the target behavior determine the set of activities that can be considered concrete indicators of change. Once these have been adopted, a person can be said to be in the action phase. For example, experience with people trying to quit smoking has shown that nothing less than 100% cessation can be considered action. Dropping back from two packs a day down to half a pack is a positive step, but does not constitute action for the purposes of TTM. People in the action phase report higher levels of temptation, and that it requires more attention and energy to continue the newly adopted behavior.

- **Maintenance**

The maintenance phase is characterized by a marked drop in reported levels of temptation, an increase in feelings of self-efficacy, and a continued positive decisional balance. These measures will be described in more detail below.

- **Termination/Acquisition**

The distinction between the maintenance and acquisition phases is less clear, and for some behaviors it may not ever be possible for a person to be completely free of temptation to revert or relapse into the unhealthy behavior that was changed.

An attractive aspect of these six stages is that there is a concrete operational definition for when a person is in each stage. Furthermore, there is a relatively straightforward strategy for developing interventions designed to help a person move forward at each stage. TTM describes ten processes associated with lasting change that can be used in a successful intervention strategy:

- **Consciousness-Raising**

These interventions are designed to provide concrete facts and figures related to the behavior in question. The individual will learn more about the pros of changing and the cons associated with not changing one's behavior.

- **Dramatic-Relief**

This type of intervention is primarily emotional and may take the form of testimonials, pep-rallies, or other sorts of events or experiences designed to create a hopeful and positive attitude toward change, and a belief that change is possible.

- **Self-Reevaluation**

Journals or diaries, therapy, and counseling are all intervention strategies that involve some measure of introspection and examination of one's current behavior. The goal is to take an honest look at current problem behaviors so that there is a clear understanding of where one should go next.

- **Environmental-Reevaluation**

This involves an assessment of how one's behavior impacts the people and places one inhabits. It involves recognition that one's actions can serve as an influence or role model to others. It may involve rational or emotional processes.

- **Self-Liberation**

Self-Liberation is associated with what people refer to as “willpower” and actions toward self-liberation may take the form of promises or commitments, e.g. New Year’s resolutions, to make certain changes in one’s behavior. The chances of success increase when multiple options (three options seems optimal) for change are available. The perception of choice seems to make it easier for people to choose to keep their commitments.

- **Social-Liberation**

This process involves increased opportunities in social environments, such as smoke-free zones, salad bars, and other places where healthy behavior is socially promoted.

- **Counterconditioning**

Counterconditioning involves learning behaviors that can serve as a substitute for the unhealthy behaviors being replaced. Chewing gum instead of smoking cigarettes is an example.

- **Stimulus-Control**

This process involves examining and removing temptations from the immediate surroundings. If temptations, e.g. cigarettes or unhealthy foods, cannot be removed, it may be necessary for the person to alter his/her environment by moving to a new location or taking a new route. Taking part in self-help groups or otherwise putting oneself into a positive environment also fall into this category.

- **Contingency-Management**

When people engaged in behavioral change institute rewards and/or punishments in reaction to their efforts to change, they are using contingency management.

- **Helping-Relationships**

This involves enlisting the aid and support of close friends, family members, or advisors who can be supportive throughout the stages of change.

Application of the various processes has been found to be appropriate during different stages. A key question when deciding how to help a person decide how best to change their behavior is exactly when each of the processes or treatments above should be applied? Three concrete measures for identifying in what stage a person resides have been developed, which are decisional balance, self-efficacy, and situational temptation. Next each of the measures and how it contributes to making decisions about selecting an intervention strategy will be described.

Decisional balance refers to the subject's self-reported perception of the pros and cons associated with making a given behavioral change. It is a fairly simple measure—a person is simply asked to list as many pros and cons as he or she can think of related to a given behavioral change. The measure is based upon Janis and Mann's work which broke pros and cons down into eight categories (Janis and Mann, 1977). Velicer et al. (1985) found, however, that the eight categories were overly complex and not as robust as merely recording pros and cons in two categories. Over a study of 12 health-related behaviors it was found that in every case movement towards positive change was correlated with a significant increase in the perception of pros coupled with a decrease in the perception of cons related to change (Prochaska, et al., 1994). This last study found that, on average, behavioral change was accompanied by a one standard deviation increase in the perception of the pros and a half standard deviation decrease in the perception of cons.

Self-efficacy and *temptation* are the second and third measures of stage of change in TTM. The measure of self-efficacy was adopted from Bandura (1982), and is “the situation-specific confidence people have that they can cope with high risk situations without relapsing to their unhealthy or high risk habit,” (Prochaska and Velicer, 1997, p40). Temptation is a measure of the degree of presence of factors such as emotional distress, positive social atmosphere, and craving. Questionnaires measuring these two constructs are developed with respect to the target behaviors. Positive change and the ability to maintain a target behavior are associated with an increase in self-efficacy and a decrease in the experience of temptation over time. Based upon these measures it is possible to design individualized interventions that have shown to be very effective at achieving favorable measures of change (Prochaska and Velicer, 1997):

- *Recruitment*—the proportion of the number of people contacted to the number of people who enroll into a behavioral change intervention program
- *Retention*—measured at fixed intervals (e.g. every three months), a comparative measure showing attrition rates between TTM-based and other intervention programs
- *Progress*—a measure of how much time people spend in each stage
- *Process*—participants’ reported satisfaction with the intervention process
- *Outcomes*—a measure of the proportion of participants that make it to the maintenance stage

The transtheoretical model has a number of aspects that make it appealing as a framework for attacking the problem of changing behaviors—in this context, teachers’ classroom behaviors. First of all, it is general enough to be applied to most any type of behavior. The authors of the model have developed interventions to address smoking and

other drug use, weight control, sun exposure, safer sex practice, and having regular mammograms (Cancer Prevention Resource Center). Second, the principles of the model appear to be applicable to organizations or groups (Prochaska, et al., 2001a). Third, and very importantly, the model gives specific guidelines for measuring and evaluating the progress of the intervention at each stage which makes it attractive in situations when accountability is an issue, such as in educational reform. Fourth, the developers of TTM have documented examples of how to set up and run a large-scale intervention based on their strategies and have even shown that the model is amenable to automation via expert systems (Velicer and Prochaska, 1999). Lastly, and most relevant to the current discussion, is that the principles of TTM are in accord with the research findings describing successful teacher professional development programs (Fishman, et al., 2003; Richardson, 2003b).

2.4.6 The Fit Between TTM and KBA

Combining aspects of the transtheoretical model and the KBA model could be a very powerful tool guiding the implementation of and research about developing teachers' professional competence. On the one hand, while TTM is a well-developed integrative framework which can guide the implementation of large scale behavioral change efforts, including baseline, intermediate, and outcome metrics, TTM can only guide these efforts at high level because TTM lacks understanding of the subject-matter specific to any one particular behavior. The TTM approach has to be developed and tailored to each new target behavior including the development of new metrics for decisional balance, self-efficacy and temptation, as well as developing a repertoire of interventions that fall into the various process categories such as awareness raising and helping relationships. The

KBA framework can fill this need because it is based upon decades of research into how best to develop teachers' competencies. The research upon which KBA is based can be harnessed to create the metrics needed to implement teacher PD following a TTM framework. In other words, KBA has filled in the specific knowledge necessary to begin a large-scale TTM-based intervention with teachers.

On the other hand, the KBA framework is hampered at this point by a lack of scalability. The methods employed by KBA researchers are labor intensive and don't lend easily to large-scale implementations as would be necessary for a school-system or nation-wide teacher reform project. TTM fills this gap by providing not only a concrete set of measures, but also guidance for the timing of implementation, guidance on how to recruit and retain trainees, and how to sequence the timing of the introduction of the different intervention processes. TTM-based interventions have been shown to be scalable, as in (Prochaska, et al., 2001b), which successfully recruited and managed the interventions with 4144 Rhode Island smokers, representing 80% of people phoned using a random-digit dialing method. KBA researchers still discuss the processes that facilitate and enable teacher change in general terms. TTM offers a well-defined set of processes that can be adopted purposively within a KBA-driven framework.

2.5 Teachers' Competence at Assessment

Having established that teacher effectiveness is important to student learning, and having explored current research with respect to teacher development, it is time to turn to the issue of what aspect of teacher practice is most in need of remediation. Teachers' competence at assessment has been chosen for this analysis. The literature on teachers'

competence at assessment reviewed below motivates this choice. This section will show that teachers are not well-trained at doing assessment. Next, two broad categories of assessment will be differentiated—summative and formative—and the literature demonstrating that formative assessment is much better in terms of fostering student learning will be covered. This section will briefly describe some good formative techniques and end with a discussion of how it is possible to measure formative practice in teachers. The discussion will lead back to the KBA framework described above, but first, a return to Bloom's work on mastery learning is in order, to focus on the role of formative assessment in improving learning.

As noted in Bloom's (1984) study, there was a remarkable increase in learning between students who were taught in traditional conditions versus students who were taught under the mastery learning condition. A key difference between traditional and mastery learning is the use of classroom assessments to help students understand their own learning progress. This section examines literature on formative assessment and teachers' competence at assessment.

Teachers' competence at assessment leaves much to be desired. Stiggins (2001) expressed this view in an invited address to the National Council on Measurement in Education (NCME). Stiggins et al. (1989) performed an extensive ethnographic study of high school teachers' assessment practices and found that teachers commonly are responsible for daily assessment of 150 students or more, must develop, collect, and report on assessments for a wide range of consumers—students, parents, administrators—rarely understand how to align assessment with the learning targets that were set for the students, and, unfortunately, often do not have mastery over the subjects they are

teaching. Stiggins also found that most states do not have strict requirements for teachers to understand both the theory and practice of effective classroom assessment, nor do they require this knowledge in their school administrators.

Terry Crooks did a meta-analysis of over 200 studies of the use of assessments in classrooms to show how they can be used to further learning and guide students' decision-making with regard to what, how often, and how much they study (Crooks, 1988). Crooks found that students take implicit cues from instructors that the material found in tests must be the material that is most valuable to learn, since it is the material upon which they are being evaluated. When the content of the tests is compared to the stated goals for a course or unit, there is likely to be little alignment, with an over-emphasis on rote knowledge. Crooks found that teachers were poorly trained at the use of assessments. Other research has also found that teachers are not good at creating tests (Benton, et al., 2004; Carter, 1984; Fleming and Chambers, 1983), aren't good at assessing the quality of tests designed by others (Carter, 1984), may not deliver them often enough (Bangert-Drowns, et al., 1991), and don't provide appropriate feedback to students based on test results (Brookhart, 1993; Stiggins, et al., 1989; Whitmer, 1983). Specifically with reference to the questions that teachers ask, either on a test or verbally, studies find an overemphasis on questions that assess only rote learning or memory, and do not tap higher-orders of thinking (Black and Wiliam, 1998b). Yet higher-order thinking is exactly what should be promoted (National Council of Teachers of Mathematics, 2000; National Research Council, 1996).

Black et al. (2003) advocate a form of assessment known as "assessment for learning" or "formative assessment." Formative assessment exists as the counterpart to

summative assessment, which Bloom, Hastings, and Madaus (Bloom, et al., 1971) defined as those tests given at the end of episodes of teaching (units, courses, etc.) for the purpose of grading or certifying students, or for evaluating the effectiveness of a curriculum (p117). Summative assessment is the main type of assessment used in the “traditional” classroom as described in Bloom’s (1984) study. Summative assessment tends to serve actors outside of the classroom, i.e., school administrators, parents, and legislators, and only indirectly serves students by way of teacher oversight, and resource allocation at the school and district level. Formative assessment, on the other hand, is designed to serve the needs of learners directly by providing immediate, constructive, and understandable feedback on performance or understanding. Summative assessment generally results in scores, letter, or number grades and has been shown to have either a neutral, or even negative effect on learning, even if accompanied by more constructive feedback (Butler, 1988; Deci, et al., 2001). Formative assessment, on the other hand, generally results in qualitative feedback that provides specific, constructive, task-related guidance. Black and Wiliam (1998b) arrived at their conclusion regarding the superiority of formative assessment after considering over 600 empirical studies on the impacts of various forms of educational assessment.

Black et al. (2003) cite Sadler (1989) when they say that:

...the core of the activity of formative assessment lies in the sequence of two actions. The first is the perception by the learner of a gap between a desired goal and his or her present state (of knowledge and/or understanding and/or skill). The second is the action taken by the learner to close that gap to attain the desired goal (Black et al., 2003, p14).

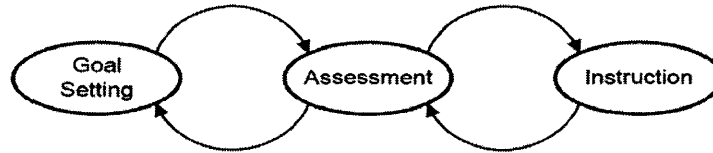


Figure 2.2—The Role of Assessment in Effective Instruction

It is very important to note that the primary actor in the formative assessment scenario described above is the learner. It is the learner who has the “desired goal,” and it is the learner who takes action to bridge the gap once he or she has, through some form of assessment, recognized that he or she hasn’t yet reached the goal. Teachers are not mentioned at all in this scenario, and ostensibly play the roles of helping in goal setting, designing and administering assessments, fostering self-assessment, and providing instruction when necessary to bridge the gap between goals and the current state. A visualization of this process can be seen in Figure 2.2. This figure places assessment at the center of the learning process. Assessment is necessary to set a baseline from which learning goals can be established. Instruction then can guide students in their progress toward the goal. Regular assessment helps students monitor their own progress through the learning process.

This is not to say that Black et al. and Sadler discount the role of the teacher. In fact, their writings mainly describe how teachers can work with students to foster a sense of personal curiosity, personal responsibility, and personal initiative in their own learning activities. Black et al.’s (2003) account is of teachers working to develop formative practices in their classrooms. A great deal of their motivation in taking on this project is the above cited finding that in general, teachers are not very competent at assessment.

In a review of empirical studies of assessment practices Black and Wiliam (1998a) find that for teaching interventions which attempt to institute formative assessment practices, the average effect size of learning gains is on the order of 0.4 to 0.7.

They state the practical implications of these numbers as follows:

An effect size of 0.4 would mean that the average pupil involved in an innovation would record the same achievement as a pupil in the top 35% of those not so involved.

An effect size gain of 0.7 in the recent international comparative studies in mathematics would have raised the score of a nation in the middle of the pack of 41 countries (e.g., the U.S.) to one of the top five. (Black and Wiliam, 1998a, p141)

2.5.1 Types of Practices Exemplifying Formative Assessment

In this next section three examples of formative assessment practice that can help improve student learning will be examined. The first practice is that of teachers asking “higher order” questions. The second practice is when teachers provide high-quality, useful feedback on assignments to their students. The third practice is when teachers encourage students to engage in self-assessment. After discussing these three examples, the logistical question of how the adoption of such practices can be measured and studied will be addressed.

2.5.1.1 Higher-Order Questions

When teachers ask higher-order questions, students learn more. By higher-order is meant questions that ask students to do more than recall information from memory, i.e. such items as facts, figures, and definitions (Anderson and Krathwohl, 2001). In one recent study it was observed that students rarely learned unless they reached an impasse—a situation when the student gets stuck in answering a question, detects an error, or answers correctly but expresses uncertainty about why the answer is correct (VanLehn, et al.,

2003). Effective teaching might then be associated with asking students questions that bring them to an impasse—an impasse that can only be broken through learning. Furthermore, in order to get students to do higher-order learning in line with national standards in math and science (National Council of Teachers of Mathematics, 2000; National Research Council, 1996) it follows that breaking the impasse must require students to apply concepts, evaluate options, and think deeply.

Unfortunately, a majority of the questions that teachers ask do not tap higher-order thinking (Stiggins, et al., 1989), and their methods of asking questions in the classroom do not allow students sufficient time to formulate a reasoned response (Rowe, 1974). A majority of the questions that teachers ask in classrooms assess nothing other than how well students read and remember facts in a textbook, or the previous day's lesson. The questions are not designed to elicit the depth of students' knowledge or understanding on the topics being covered. Furthermore, teachers tend to solicit an immediate response from a select set of students within any given class who typically respond to questions (Black, et al., 2003). In contrast, using a "hands down" strategy which forces all students to participate, and increasing the "wait time" from asking a question to soliciting an answer to as little as five seconds can produce remarkable improvement in student engagement and responsiveness.

Formative assessment supports using well-designed questions not only to encourage deeper thinking, but also to diagnose student misconceptions. A recent study showed that even in one-on-one tutoring sessions, instructors are not very good at monitoring the subjective understanding of their students (Chi, et al., 2004), although this did not seem to hamper learning in such individualized settings (Chi, et al., 2001). Of

course the major problem is logistic—how does a classroom teacher, who sees upwards of 100 students per day, monitor and respond individually to all of them? The DIAGNOSER project built a system to allow high school physics teachers to do just that (Thissen-Roe, et al., 2004), built upon the concept of facet-based instruction (Minstrell, 1992). Facets are conceptions of the world, or procedures for problem-solving held by students. In any given domain some facets are “correct” interpretations of the world, and others are less so. The role of a teacher in facet-based instruction is to determine which facets are held by students and convince them to replace these with the target facet, or learning goal.

A major bottleneck in the development of systems such as the DIAGNOSER or Cognitive Tutor™ (Blessing, 2003; Koedinger, 1998) is that they are specialized and depend on a detailed cognitive task analysis of the domain being taught. While for some subjects that have received extensive study such as Algebra I, or Newtonian motion, empirical analyses of students’ thinking are available, but there are still a large number of subject areas that have not received such attention. It will be some time before such systems based on deep, empirically confirmed understanding of student cognition will be able to be developed, if ever. An alternative system is proposed that encourages teachers to develop such deep understanding on their own. Encouraging teachers to develop such knowledge is a part of what Shulman (1986) referred to as pedagogical content knowledge. Recent depictions of the teacher as a learner have emphasized building up such knowledge (Iszák and Sherin, 2003; Sherin, 2002).

2.5.1.2 Providing Useful Feedback

A deep understanding of students' thinking gained from asking higher-order questions will allow teachers to develop in the second practice of formative assessment—providing useful feedback. The quality of feedback has been shown to have one of the largest impacts on students' learning (Bangert-Drowns, et al., 1991). A problem is that if teachers provide students with a number or letter grade, students will ignore any other written comments on their work, although it is precisely these comments which would help them the most. This has led to the successful implementation of “comment-only” marking as a formative practice (Black, et al., 2003). Effective feedback does three things: points out to the student what he or she did well and what things need work, reminds the student of the learning goal, and gives actionable advice on how to bridge the gap. The obvious problem is that providing such high-quality feedback is extremely time-consuming. Software systems seem a likely solution.

Systems such as the DIAGNOSER facilitate the delivery of individualized feedback based on test performance (Thissen-Roe, et al., 2004). Commercial systems, e.g. WebCT™, support tying feedback to specific responses, but do not offer guidance as to what type of feedback teachers should provide. The Cognitive Tutor™ provides feedback in real time while students are in the act of problem solving to help them recover from errors or overcome roadblocks in their understanding. Expert systems have been used successfully in healthcare settings to provide tailored instructions and feedback to people for taking medication and changing unhealthy behaviors such as smoking (Hirst, et al., 1997; Velicer and Prochaska, 1999). These expert systems allow healthcare professionals to anticipate the kinds of issues patients will face and prepare their

responses ahead of time. An interesting feature of the expert healthcare systems is that the output is printed, which means that patients do not need to interact with a computer directly. This is important for educational settings where there is not a computer for every student in every class and teachers must give feedback orally or on paper.

Providing high-quality feedback to students may not be enough, in and of itself—students must be taught how to take advantage of such feedback. While it may seem apparent to the instructor how a student should act upon reading detailed comments on a piece of work, students have been conditioned through years of summative feedback to be focused on and motivated by gold stars, letter grades, and number grades. Upon moving to a comment-only marking style, a teacher will receive many questions of the “Yes, I read the comments, but what did I really get?” variety from students. Such moments are an opportune time to refocus students’ attention on the intrinsic goals of the curriculum and the value of the content and skills being studied.

Once accustomed to this rich feedback, students will not want to go back. To illustrate this, (Black, et al., 2003, p67) recount an incident that occurred in a class of students that had become accustomed to rich, formative feedback. They relate that:

One class, who were subsequently taught by a teacher not emphasizing assessment for learning, surprised that teacher by complaining: ‘Look, we’ve told you we don’t understand this. Why are you going on to the next topic?’

2.5.1.3 Encouraging Students to Engage in Self-Assessment

Self-assessment is the third formative practice to be discussed. Feedback is not complete unless students take advantage of it to improve their learning. Toward this goal it is important that students be taught to assess their own progress. Self-assessment and peer assessment, if used appropriately, encourage students to think deeply about success

criteria and then review their own performances as measured against these criteria. In addition, if students are involved in their own performance assessment, it can reduce the amount of time that teachers need to spend marking assessments (Black and Wiliam, 1998b). This will free teachers to spend time providing detailed feedback on a smaller number of key assignments. In general, self-assessment practices fall into the category of meta-cognitive knowledge (Anderson and Krathwohl, 2001), and aid students in become self-sufficient, lifelong learners.

A problem with these three practices—higher-order questions, high quality feedback, and teaching self-assessment—is that a broad range of practices that fit into those categories have been shown to be successful, making it hard to measure the effect of adopting formative assessment techniques. Judgments must be made as to whether or not a given behavior represents a formative approach to assessment, and indeed, the same practice can be either summative or formative depending on the context. Luckily, however, as discussed earlier with respect to the KBA framework, there is a relatively stable set of attitudes and beliefs associated with the adoption of formative practice. It is likely that measuring these can provide insight into teachers' adoption of formative practice.

2.5.2 Attitudes and Beliefs Associated with Formative Assessment Practice

While the specific practices that teachers take on when they adopt formative assessment will differ substantially from teacher to teacher, their attitudes and beliefs change in a rather uniform way. It is generally agreed that a teacher's number one priority should be the learning of his or her students (Sherin, 2002), yet teachers' behavior often indicates that this is not the case. The literature on teacher stress offers many possible explanations

as to what sorts of pressures may lead teachers to put priorities other than learning ahead (Kyriacou, 2001). An oft-cited example is the pressure put on teachers to have their students perform well on standardized tests. Such pressure leads them to drill students on question types that are representative of those on the test, but which may not be the best way to foster understanding. Additional time may be spent on things like guessing strategies, and other techniques designed to improve scores without affecting learning. Regardless of these pressures, it was found that teachers who successfully adopted formative assessment practices had a renewed and strengthened belief that student learning should be the number one priority in the classroom (Black, et al., 2003). These teachers were able to put aside their fears of not “covering the material,” able to avoid the temptation to “teach to the test,” and were able to focus on students’ understanding. (This is also an example of the TTM principle of teachers who have reached the maintenance phase of change being able to resist situational temptation.)

In conjunction with their renewed focus on learning, teachers’ expectations of their students changed. The connection between teacher expectations and student performance is well-established (Brophy, 1983). Whereas at the beginning of the study the teachers tended to see student ability as relatively fixed, teachers began to believe that student ability was fluid. Teachers who had begun to assume all of the responsibility for their students’ learning, began to shift that responsibility back to the students. Armed with new formative teaching practices, the students were ready to take the responsibility for their own learning.

Based on the above, it is believed that attempting to develop teachers’ professional competence in the use of formative assessment is a goal likely to produce

significant results. However, assessment takes many forms and therefore it is advisable to narrow the focus one more degree before turning to the issue of how to design an information system to support the development of teachers' competence at assessment.

2.6 Multiple-Choice Questions

The chosen focus for assessment practice is the development of competence in the writing and use of selected response items, more commonly known as multiple-choice questions (MCQs). The motivation for this choice is as follows:

4. Multiple-choice questions are frequently used because they simplify grading. Yet, they are hard to construct (McKeachie, 2002).
5. Students see them on standardized tests, and hence getting practice in classroom settings is desirable.
6. From a programming perspective, multiple-choice items lend themselves well to online delivery and grading, particularly for real-time use with simple, and inexpensive wireless technology becoming more common in classrooms (see the section cataloguing technology for assessment below).
7. There is a great deal of collected wisdom as to how to create high-quality multiple-choice questions (Haladyna, et al., 2002). Much of this wisdom could easily be implemented in software.
8. MCQs are frequently distributed with textbooks in a format ready for import into the more common course management software. These questions appear to vary widely in quality, and it is likely given the literature on assessment

competence that instructors will not accurately be able to assess this quality, and/or may ignore issues of quality for the sake of convenience.

9. MCQs offer the opportunity of instantaneous online feedback for students.

As demonstrated above, teachers are inadequately skilled in the development of assessment items in general, and MCQs, in particular, are no exception (Haladyna, et al., 2002). In professional test development environments, a formalized process exists for developing test items, an example of which is shown in Figure 2.3. In practice, professional test development may involve additional rounds of item evaluation such as fairness review (Hambleton and Rogers, 1995) and difficulty evaluation employing a variety of techniques, such as expert evaluation (Prestwood and Weiss, 1977), test-taker evaluation (Newman and Taube, 1996), feature-based evaluation (Green, 1983), and statistical evaluation based on classical test theory or item-response theory (IRT) (Hambleton, et al., 1991). For high-stakes testing the discard rate for items developed in such a process may be as high as 40% (Haladyna, 2001). For the purposes of classroom-based, formative assessment, such a rigorous process is probably not necessary, but it is a good starting point from which to identify what types of “best practices” should be targeted for inclusion in a teacher professional development program focusing on formative use of MCQs.

Haladyna’s question development process (Figure 2.3) begins with identifying the purpose and content of the test to be developed. Given the earlier description of the importance of goal-setting in a formative learning environment, these steps in the MCQ development process should become a part of a classroom teacher’s question-writing process as well. Haladyna’s framework for doing this is very similar to that of the

1. Identify the purpose of the test
2. Identify the content to be tested
3. Develop a set of test specifications
4. Identify the MC item formats to be used
5. Identify the item-writing guidelines to be followed
6. Establish the number of items needed for the item bank
7. Identify and recruit subject-matter experts (SMEs) to write the items
8. Give SMEs an item writing guide
9. Provide item-writing training to SMEs
10. Give SMEs specific item-writing assignments
11. Have one or two other SMEs review each new item
12. Place new items in an electronic file (item bank)
13. Administer these items to a group (field test)
14. Evaluate performance and Retain, Revise or Retire each item

Figure 2.3—Steps in Professional Item Development (Haladyna, 2001, p4)

revised Bloom's taxonomy of educational objectives (Anderson and Krathwohl, 2001). In this taxonomy (Figure 2.4) objectives are classified along two dimensions—cognitive and knowledge content. The levels of the cognitive dimension are ordered from least complex to most complex, and a major assumption of the taxonomy is that higher cognitive levels subsume lower levels, so that for example, a student who has mastered application of a given piece of knowledge is assumed to have also mastered recall and understanding of that knowledge. So far research has provided only mixed support for the validity of this assumption (Anderson and Krathwohl, 2001, see Ch. 16), but in practical situations the lower levels of the framework provide useful guidance for test item developers, as is indicated by the shaded area in Figure 2.4. The knowledge dimension of the framework is relatively new and has not yet received much empirical testing; yet it, too, appears to be useful for item development. The practical implications of this are discussed presently.

There are several problems that arise when educational objectives are not adequately specified for a given assessment. One problem is a bias towards measuring

only knowledge that is easy to measure, a problem sometimes referred to in scientific research as the drunkard's paradox². This problem manifests itself in questions that when analyzed appear only to assess a student's memory of facts or potentially trivial information. These questions fail to address the underlying significance or conceptual importance of the knowledge the student is being asked to recall. This is not to say that there are not cases when it is appropriate to use questions assessing memory; however if the objectives of an assessment are not planned out in advance there is a tendency for MCQs to over-emphasize rote memory to the detriment of higher order cognition. A second problem when objectives are not clearly defined is that tests fail to cover the entire range of subjects that were taught or introduced in a unit of content. Instead, tests over-emphasize only a subset of the concepts that should really be covered. These problems have important consequences because students receive cues about what is important to learn and spend their time on, from the items of the tests that evaluate them (Black and Wiliam, 1998b; Crooks, 1988).

The Knowledge Dimension	The Cognitive Process Dimension					
	Remember	Understand	Apply	Analyze	Evaluate	Create
Factual Knowledge						
Conceptual Knowledge						
Procedural Knowledge						
Meta-Cognitive Knowledge						

Figure 2.4—Revised Bloom's Taxonomy of Educational Objectives (Anderson and Krathwohl, 2001). Shaded region indicates areas Haladyna (2001) considers appropriate for assessment with MCQs.

² As the story goes, a drunk is crawling around on his hands and knees under a streetlamp when a second drunk comes over and asks him what he's doing. "Looking for my car keys," he replies at which point the second drunk graciously offers to help him look. After ten or fifteen minutes of both drunks crawling around under the lamp, the second drunk asks, "Are you sure you lost your keys here?" In response to which the first drunk shakes his head, points off in the distance and says, "No. I lost them over there, but the light is better here."

Content Concerns	Writing the Choices
1. Specific content, single specific mental behavior	18. As many effective choices as possible
2. Important content; avoids trivial content	19. Only ONE right answer
3. Novel material for testing higher level learning	20. Location of right answer is varied
4. Content independent of other items	21. Choices in logical or numeric order
5. Not too specific, nor too general	22. Choices are independent; non-overlapping
6. Not based on opinion	23. Homogenous content/grammatical structure
7. Not a trick question	24. Choices are about the same length
8. Vocabulary simple for group being tested	25. <i>None-of-the-above</i> used very carefully
Formatting Concerns	26. Avoids all-of-the-above
9. MC, TF, MTF, matching; NOT complex	27. Choices phrased positively;
10. Vertical formatting	28. Avoids giving away right answer by using:
Style Concerns	28a. Specific determiners, e.g. always, never
11. Edited and proofed	28b. Clang associations
12. Correct grammar, punctuation, caps, and spelling	28c. Grammatical inconsistency clues
13. Minimize reading for each item	28d. Conspicuous correct choices
Writing the Stem	28e. Pairs or triplets of options
14. Directions in the stem are clear	28f. Blatantly absurd or ridiculous options
15. Central idea is in the stem, not in the choices	29. Make all distractors plausible
16. Avoids excessive verbiage	30. Uses typical errors of students as distractors
17. Avoids or cautiously uses NOT or EXCEPT	31. Cautious use of humor

Figure 2.5—Guidelines for MCQs (Haladyna, et al., 2002)

Once the educational goals for a unit of instruction have been established, it becomes possible to begin creating items to assess when students have reached these objectives. A number of researchers have done studies to determine how best to construct MCQs so that they will have the best chance of providing valid measures of target knowledge. Haladyna et al.'s (Haladyna, et al., 2002) review of the literature compiled a list of 31 guidelines for writing MCQs (Figure 2.5), and summarized the empirical literature validating these guidelines. The process outlined in Figure 2.3 indicates that it is necessary to identify and train subject-matter experts (SMEs) in order to develop the questions. In a classroom setting, the teacher, and perhaps colleagues in the same department, *are* the SMEs. Setting aside Stiggins' (1989) observation that too often teachers are not masters of the subjects they are asked to teach, the professional development issue then becomes, how can the following be accomplished:

1. Train teachers to write good MCQs

2. Train teachers to understand and interpret the results of MCQ-based tests
3. Train teachers how to use MCQs as a formative assessment method
4. Convince them to adopt the habit of doing these things

The assumption is that if teachers begin to write good MCQs and use them formatively on a regular basis, student learning will improve significantly. In answer to the first goal, Haladyna (2001) has written a very accessible book which goes step-by-step through the process of how to develop a high-quality MCQ. To answer the other three goals, a system design will be proposed which incorporates both the research on learning, behavioral change, and teacher professional development reviewed above as well as research on the design of persuasive systems and technology's tendency to shape problems. The literature on structuration, how technology shapes problems, is reviewed below.

2.7 Discussion of the Assessment/PD Literature—Implications for System Design

So far this review has spent little, if any, time discussing the role of information systems in the process of developing and using MCQs for formative assessment. The rest of the review aims to do just this. In this section, the research described so far will be used to build a foundation upon which a system design will eventually be proposed. The next section will follow accepted systems development practice and survey extant systems that have similar and/or related functionality, making note of features that should or should not be incorporated into the final system design. Two more sections will follow which bring information systems theories to bear on the design problem, namely theories on persuasive technology and on the role of information systems in shaping problems and

how they are solved. The final section of this paper will revisit the framework to be developed in this section and discuss the opportunities to support the teacher adoption of formative use of MCQs with technology.

2.8 Catalogue of Extant Systems

A standard practice in software development is to study similar and related systems that have already been built to help identify what features will be critical to the system under consideration. Examining the software that has already been built will highlight trends, strengths, weaknesses, and opportunities for a new system. There is a broad and growing range of software available for teachers. There are many types of software to support learning and teaching in general. This section will limit its discussion specifically to systems which support some form of assessment of students.

2.8.1 Commercial Systems

Two major systems most frequently used in higher education are WebCT and Blackboard. These two systems are very large and must be licensed at the university level because of their cost and the complexity of deployment. Both provide a comprehensive course management system and claim to be compliant with standards such as IMS and SCORM. Tools to support assessment are incorporated into both systems. WebCT offers instructors a question database into which questions can either be entered manually or uploaded. It has become more and more common for textbook publishers to distribute question sets along with textbooks that are ready to be uploaded into these packages. Once questions have been entered into the database, the instructor can select questions to form the basis of tests that can then be delivered online to students.

A range of test parameters can be controlled by the teacher such as the number of questions, the order in which they are delivered, the number of times a student may take a test, the time allowed, whether or not students must answer questions sequentially or can revisit skipped items, and also the specific IP addresses from which the test may be taken (to prevent cheating?). Besides MCQs there are templates for essay, matching, calculated (mathematical) and short answer questions. Once students take a test, there are options which support automatic scoring. Although essay questions must be scored manually, technologically advanced teachers can enter regular expressions to create a scoring guide for the short answer questions. Scoring other question types is relatively straightforward. Some other advanced features of WebCT include the ability to create question sets—groups of roughly equivalent questions from which one or more will be randomly selected for inclusion on any given student’s test—meaning that different students will get different, randomly generated tests. WebCT does not have the same level of integrity checking controls that a system such as TestNav (described next) does. Finally, WebCT provides some statistical analysis of questions and tests. The proportion of students who answered a given question correctly is broken down by the percentage of students in the upper lower quartiles, as well as overall, who got the question correct. A discrimination score is calculated to indicate the degree to which certain questions separate “high” and “low” ability test-takers. Given the literature on teachers’ understanding of psychometrics, it is not likely that many will be able to make effective use of these statistics.

TestNav is the test delivery component of the Progress Assessment Series™ (PASeries) assessments sold by Pearson Education. When school systems purchase or

subscribe to use the PASeries tools they are given access to a large bank of test items that are “scientifically based” and generate achievement or proficiency scores based on two proprietary scales called Lexiles and Quantiles. Scores are designed to cover the range of ability from primary through secondary education so that scores can be compared and progress monitored from year to year, consistent with a value-added approach to learning measurement. Teachers may select the questions that get delivered via TestNav but may not be able to create their own questions. TestNav, itself, offers some innovations to test takers including the ability to cross out answer choices that are considered to be incorrect³, the use of tools like a geometric compass, and the ability to highlight text in a given question. Teachers may specify what tools are available when creating tests. When the TestNav program is launched it takes up the entire computer screen. Test-takers are instructed that if they close the screen—i.e. if they try to open up, say, a web browser to look up the answer or chat with a friend who knows the answer—that the test will end immediately. Combined with strict time limits, question randomization and other techniques, TestNav is likely to have one of the best feature sets for preventing cheating among currently available tools.

Pearson’s TestNav is only one of several standards-based diagnostic assessment suites offered by major industry players. CTB/McGraw Hill sells a system called i-know that offers “fast, reliable diagnostic information on student skills to target teaching strategies” (<http://www.ctb.com>). Some of these systems offer tools to teachers that will use test results to create groups of students who either all appear to have the same

³ This is consistent with a test-taking strategy for MCQs which advocates guessing if a test taker can eliminate at least one of the choices from consideration. Statistically it has been shown that on high-stakes tests, the chances of guessing correctly go up tremendously if students can eliminate at least one of the distractors. It should be noted that such a strategy contributes little to learning.

misconceptions so that the teacher can spend extra time with them on the specific issue, or groups of students made up of those who did and did not show a misconception to allow for learning from peers. These systems tend to be quite expensive, and are sold mainly to entire schools, school districts, or even states. Staff development and training is generally sold as part of the package.

2.8.2 Systems Generated from Basic Research

The Cognitive Tutor (Blessing, 2003; Koedinger, 1998) system was developed at Carnegie Mellon and the University of Pittsburgh, and has shown some of the best potential for a software system that promotes learning. The Cognitive Tutor is designed for students to use on their own, i.e. without a human instructor, and allows students to progress through the steps of a problem, asking for hints if they need it, and repeating problems until a skill has been learned. If students make a mistake, the system helps them reason their way out of it. The dialogue between student and system is based on a detailed cognitive task analysis of the specific learning task, e.g. two-digit addition. Cognitive Tutor embeds a mastery learning approach to learning and has been able to achieve 2-sigma learning gains in students learning algebra. This is reminiscent of Bloom's study. One of the major drawbacks of the system is that performing a cognitive task analysis of every skill important for students to learn and then coding these analyses into the software is an enormous task. Not surprisingly, the system has had more early successes in subject areas such as algebra, which are more amenable to software implementation, than it has had in verbal subject areas such as reading. While this system appears fantastic for students, it does little if anything to help teachers better understand their students and the mistakes that they make.

One feature of Cognitive Tutor is shared by another system called CAPA (Thoennessen, et al., 1999), which supports self-paced learning of physics. The CAPA system bases each problem on a physical model for which the parameters are generated randomly as the student uses it, meaning that the student can receive as many practice problems to work as necessary until a concept is mastered. The concept of item models (Bejar, et al., 2003) is more formally applied in another piece of software under development at the Educational Testing Service called the Math Test Creation Assistant (TCA) (Singley and Bennett, 2002). Item models, in theory, offer the potential to automatically generate large numbers of items with desirable psychometric qualities (for use in high-stakes assessment). In practice, however, the items generated by systems such as TCA have not produced reliable results. Such systems may be useful for low-stakes environments such as classroom practice, however.

A system which does help teachers gain insight into their students' misconceptions is the DIAGNOSER (Thissen-Roe, et al., 2004), which, as described earlier, is built upon the concept of facets (Minstrell, 1992) described earlier. Again, while there is a bottleneck in the development of systems like DIAGNOSER stemming from the sheer amount of work to be done documenting students' misconceptions about all areas of learning, this tool does follow a formative strategy in which teachers strive to learn more about individual students, and work with them to correct their misconceptions.

2.8.3 Shareware and Other Tools

Online searches for "educational software" turn up a variety of different tools from games, to drill and practice tools, to simulations, which frequently claim to be "research based." One example is a company called Gamco.com which sells a variety of tools for

teachers which help them create worksheets, make flash cards, and assess their students' learning styles among other things. Another company, GradesAndPlans.com sells a computer-based gradebook, and a word-processor-like tool for creating lesson plan templates. These types of sites tend to follow the pattern of creating electronic versions of traditional classroom learning materials. In general there does not seem to be much or any innovation and research support for the "research based" tools is spotty or missing.

2.8.4 Discussion

With the exception of the several tools highlighted above, the majority of the software designed to support teachers' role as assessors of student understanding does little, if anything to augment or improve upon the methods that teachers currently employ. There is virtually no literature supporting or rejecting the effectiveness of the tools offered by the major software vendors. These tools are undoubtedly based on research of some sort, but an observation of the school software industry is that often these tools appear to be designed more to serve the needs of the people making the purchasing decisions, i.e. school administrators, and less to serve the needs of students and teachers. Consequently, the tools emphasize the diagnostic assessment and reporting functions, which provide high-level summary data on overall student performance, which is what administrators need to demonstrate to legislators and policy-makers that their schools are performing effectively. Ironically, as was shown above, it is likely that the focus on summative assessments does little for, and may even hurt, students' learning.

Software systems that are the product of basic research, such as Cognitive Tutor, DIAGNOSER, and Math TCA, seem to show great promise in the areas that they address. The tools described here show a deep understanding of students' cognition and work to

leverage that understanding to achieve learning gains, but all suffer from the problem of scalability. Based on this observation an idea is that rather than building the expertise in every learning domain into the tools, a better approach might be to build a more general tool that leverages the expertise of teachers in their given subject matter areas. This type of tool would foster and support teacher learning (Sherin, 2002; Shulman, 1986) as an approach to improving students' learning. A major drawback of this approach is that it becomes more difficult to link learning gains in students to changes in instruction that are in turn the result of teacher learning. This issue will need to be addressed seriously in any plan to evaluate a system that aims to foster teacher learning as a means to improve student learning.

Now that a baseline has been established for the types of systems that exist to support teacher development of MCQs, theories of system building that may inform the design of an improved version of such a tool can be examined.

2.9 Persuasive Technology

One goal of the system is to influence the behavior of teachers. It is not only important that teachers understand the differences between formative and summative assessment, and important that they understand the misconceptions their students hold with respect to the subject matter being shown to them. It is also important that teachers implement, on a daily basis, the style of instruction that is fostered by this type of understanding. While the transtheoretical model and KBA provide insight into the types of interventions that may bring about this behavior and provide ways to measure the current stage of change

of the teachers being developed, a software based solution would do well to consult the literature on persuasive technology.

Captology is a term coined to refer to the notion of **Computers As Persuasive Technology**. The basic idea is that in the process of human-computer interaction computers can and do influence the behaviors that people take, that people are not always aware of this influence, and that system designers should be aware of this influence and use it ethically in the design of systems (Berdichevsky and Neuenschwander, 1999; Fogg, 1999). Given that the goal of professional development activity is the modification of teachers' knowledge, beliefs and attitudes with respect to their teaching practice, and their understanding of their students, the best system design to accomplish that would be one that persuades them to adopt the desired behavior.

The research on persuasive technologies began in the early 1990's when a research group at Stanford asked the question: "What if people's reaction to computers is fundamentally social?" (Nass, et al., 1994). They took a novel approach to this research in which they began by taking sociological and social psychological theories involving social interaction between people and replacing the word "computer" for one of the actors. Then they ran experiments collecting data to see if the theories still were accurate predictors of human behavior, even if the other actor was a computer. They also wanted to see how easy it was to induce a human to treat a computer as another human. Not only did they find that people react to computers in the same way that they react to other people, but that inducing this behavior did not require that the computer exhibit sophisticated behavior. A second study found that not only was it easy to cause this reaction, but it was also relatively easy to emulate different human personality types with

computers (Nass, et al., 1995). This was significant because it was found that people could only not recognize the personality types, but also expressed more satisfaction and perceived themselves to be more effective when the computer's "personality" matched their own.

In 1997 the group reported the first case in which they were able to document that people changed their behavior in response to either "helpful" or "unhelpful" computers (Fogg and Nass, 1997). In this study subjects first completed a task on a computer in which the computer could help them a lot, or a little. To simulate this, users performed an information search. In the "high-help" case, users received very relevant and useful results. In the "low-help" case, users received mostly irrelevant and unhelpful results. At this point, roles were reversed and the computer asked the user to help it determine the best color combination for the desktop layout. The computer showed the user a series of color combinations and asked the user to rate their attractiveness. The user could stop this at any time. As a control, half of the users performed the second task on a different, though identical computer. It turned out that users spent more time, made fewer errors, and felt more positively about their experience with the "helpful" computer as compared to the "unhelpful" one. The authors of the study felt that these results have implications for most systems that interact with users, and could be influential in, among other things, "persuading students to work longer and harder." Berdichevsky and Neuenschwander (1999) addressed the issues raised in this study by laying out eight ethical principles relating to persuasive information systems design and creating a decision-tree style procedure for making decisions based on the principles. These principles are generally what one would expect, i.e. never create technology to persuade someone to do

something that you wouldn't do yourself. However, one principle might be seen as problematic for some software developers—principle three states that creators of persuasive technology must take responsibility for “all reasonably predictable outcomes of its use.”⁴

The area of persuasive computing began to show signs of maturing with Fogg's (1998) paper defining five perspectives and research directions for the field. He formulated a definition of a “persuasive computer” as follows:

A persuasive computer is an interactive technology that...attempt[s]to shape, reinforce, or change behaviors, feelings, or thoughts about an issue, object, or action. (Fogg, 1998, p225)

This definition was further qualified by stating that since computers by themselves don't have intentions, the intention must come from the designer or adopter of the technology and could take several forms. Endogenous intention derives from the original designers of the technology. Exogenous intention derives from the providers of the technology, as in when a company provides an email program to employees to encourage them to communicate with this tool. Autogenous intention comes from the adopter of the technology as when a person buys a calorie-counting device to motivate themselves to lose weight. Fogg also defines further dimensions to the field specifying that computers typically function as tools to augment capabilities, media to provide experiences, and social actors to create relationships. Persuasive systems should be analyzed at appropriate levels ranging from individual to societal. Lastly, he identifies

⁴ This principle has been given the force of law in at least one case, that of the recently decided Grokster case where the file-sharing software developer was held liable for the illegal sharing of copyrighted material perpetrated with their tool.

some of the domains in which these systems might apply and emphasizes the need for strong ethical consideration in the development of such systems.

Clearly the ideas in the literature on captology will be applicable in the design of a system designed to convince teachers to change their current attitudes and behaviors. Since teachers do not act alone, but rather in a complex and multi-layered environment—i.e. individual lesson planning, classroom instruction, school, district, and state level interaction—it may be necessary to incorporate features into the technology that persuade actors at all levels to act in different ways. Certainly in situations involving children the ethical considerations will also take a prominent place in the system design. The software will play a social role as guide and coach. It will be a tool for creating assessments and it will also serve as a medium to transmit the created questions to students.

2.10 Technology Shapes Problems

This will be the last section of this review and bring to bear a final segment of literature on systems design that will impact the building of the tool to help teachers write better MCQs. In 1992 Orlikowski introduced a new concept which she called the structuration model of technology (Orlikowski, 1992). Up until that time, organization scientists had difficulty working out the complex relationship between organizational structure and technology. Practically speaking, organizational managers wanted to be able to predict the changes that would result in an organization because of the implementation of a particular technology. This information is crucial for selecting and designing the most appropriate technologies to further an organization's goals. Before Orlikowski's theory,

an essentially deterministic view had been taken in the sense that various theories predicted that the implementation of specific technologies would affect organizational structure and performance in more or less predictable ways. Unfortunately, reality wasn't so simple, and these deterministic models failed to be reliable tools. Orlikowski introduced the concept of a duality in which systems exist in a sort of feedback loop with organizations. In other words, technology does help shape organizations which in turn help shape technology. She advocates a line of research that would explore this relationship.

Orlikowski's model was further developed by DeSanctis and Poole (1994) who argued that the focus of analysis must be finer grained than looking only at the institutional level. Their new articulation of the model was called Adaptive Structuration Theory (AST). Because of the social relationships and work patterns within small groups within an organization AST posited that technology would be adopted and used that fit within the patterns of the group members. In turn, the more a given technology was used by the group, the more opportunities were offered for the group to change their social patterns.

In building a system to modify teachers' assessment practices, it is reasonable to predict that the technology will have some impact on the social relationships between teachers and students, between teachers and their peers, between teachers and their supervisors and others that make up the social organization of the school. Structuration and AST provide a model not only for predicting the changes, but also for designing the technology in such a way as to encourage change in the most positive direction.

Furthermore, studies based on AST will give guidance in selecting variables of interest for the evaluation and modification of the system after it is put into practice.

2.11 Discussion and Conclusion

In this section of the literature review the argument thus far will be reviewed and summarized, and then the work that was done in the dissertation will be outlined.

2.11.1 Developing a system to help teachers get better at assessment

In the introduction of this review a quotation from Leidner and Jarvenpaa's 1995 paper predicted that computer systems for education may not live up to their potential of transforming education and learning. It was shown later on in the catalog of current systems that with few exceptions, Leidner and Jarvenpaa's prognosis has been realized in the area of systems to support educational assessment. Although the role of information systems was not dealt with directly through much of this review, the central underlying question of this review has been, how can information systems improve learning?

The first approach to this problem was to examine all of the potential factors that impact learning, and then to focus on one that may have one of the greatest impacts—teacher quality. Literature examined from a macro and micro level showed that the quality of teaching has significant and potentially long-lasting effects on how much students learn. The goal of the information system then should be to help teachers improve the quality of their teaching. This led to the question: What is the best way to help teachers improve? Asking this question prompted a look at the literature on teacher professional development.

The research on teacher professional development has created a coherent set of attributes that describe a high quality PD program. Such programs are long term, develop communities of practice, meet immediate needs of participants, offer opportunities for inquiry and reflection, and use the results of past PD to inform the content of future PD. But while the characteristics of a high-quality PD program are known, selecting a set of practices to inculcate that which will best serve a large number of teachers is a more difficult challenge. Fortunately, the same research on teacher PD has uncovered the KBA framework which shows that effective teaching is predicated less on a specific set of practices than on a core set of pedagogical and content knowledge, and positive beliefs and attitudes towards one's students, and one's ability to teach those students using that pedagogy. Hence the relatively new KBA framework serves as a framework upon which to develop programs for teacher PD. The development and advancement of such research-based programs has also been called for by the president of the AERA. Reading the literature on how to encourage the incorporation of research findings into professional practice led eventually to the transtheoretical model of behavioral change, TTM. While research on TTM has focused primarily on health behaviors, the framework is formulated very generally and can apply to almost any form of human behavior. The opportunity to combine the insights of KBA with the structure offered by TTM may bring the depth of understanding of teacher behavior found in KBA to bear in a large-scale intervention program built following the guidelines of similar, successful programs built using the TTM framework. An as yet undiscussed opportunity exists to use information systems to support this intervention.

The measures developed by the TTM model have the virtue of being relatively straightforward and simple, and hence amenable to computerized delivery. It is possible to build tools for teacher support in such a way that the interfaces can not only deliver interventions, such as tutorials about particular pedagogical issues, but also gather much of the necessary feedback implicitly or explicitly. For example, questions related to a teacher's decisional balance (see section 2.4.5) could be asked when the software is launched or could be presented to teachers on screen next to task areas in a way that may elicit voluntary participation. In this way, the tools to gather data used to design future PD are built into the tools that are used to carry out the practices advocated by the PD.

Computer interfaces can be persuasive, and one goal of a system designed to improve teacher performance will be to persuade the teacher to adopt a certain set of practices related to their teaching. The principles of persuasive computing, including ethical analysis, can and should be brought to bear in the design of these teacher support tools. Furthermore, the power of the tools to shape the way that people think about education, and, in turn, for educators to think about technology will also be a factor in the system design. There may be a great deal of creative license taken with the tools which may open up new avenues to explore in the future. It will be important to be sensitive to how the tools are adopted as well as how they are not.

Finally, having an idea now of what a potential tool to support teacher development may look like, it is important to choose an area of improvement upon which to focus for the development of the first tool. The area of teachers' competence at assessment and, more specifically, teachers' ability to develop multiple-choice questions

was examined. This is an area where teachers stand to gain a lot and where tools, to date, have been relatively uninspiring. Now, the next step in this research can be discussed.

2.11.2 The Dissertation

It should be made clear that the vision outlined above for TTM-based systems to support teacher PD will take multiple years to implement and a great deal of experimentation. This review thus defines a research plan that will span the next five to ten years. For the dissertation a tool was built to help instructors write better multiple-choice questions. The chapters to follow describe how the tool was built and evaluated.

CHAPTER 3

TWO PILOT STUDIES INVESTIGATING MULTIPLE-CHOICE QUESTIONS

3.1 Introduction

Two pilot studies were performed to provide empirical support for the proposed research. The QuizViz study was conducted during the spring of 2003 and was originally part of an effort to design a better way for teachers to visualize the results of student quizzes. In the process it was discovered how poorly the questions were written and so it was decided to refocus the research on better assistance for the development of high quality questions. The results of this first study were written up and submitted as completed research (Benton, et al., 2004). The description of the study, its results and analysis are reproduced here.

The QuesGen tool was built using the wisdom gleaned from the first study along with the literature reviewed in the state of the art literature review. The second study was a dry run of the procedures that were planned for the evaluation of the tool. The evaluation plan is described in the methodology chapter of this proposal. The experiences and results of the second pilot study prompted modifications to the procedure that are described also in the methodology chapter.

This chapter is organized as follows. The first pilot is described and then the results summarized. Then the second pilot study is described and the results of that study are summarized.

3.2 The QuizViz Study

3.2.1 Description of the QuizViz Study

This report focuses on two methods that were used to analyze the questions generated by a college professor: categorization of questions into Bloom's taxonomy, and analysis of quiz results informed by a student questionnaire. Bloom's Taxonomy of Educational Objectives was first developed in the 1950's by Benjamin Bloom and his colleagues, and was an attempt to systematize the creation of test questions by categorizing the types of mental activities that were assessed by those questions (for a fuller description please refer to section 2.6 above). The original taxonomy had six categories in a single dimension, but recently, a second dimension has been added to reflect changes in the way psychometricians understand knowledge and cognition (Anderson and Krathwohl, 2001).

Twenty-four students from an upper-level undergraduate, distance-learning course in human-computer interaction (HCI) at a mid-sized, urban, northeastern, technology school participated in the study. All of the students in the course completed all course requirements online, only coming to campus for the midterm and final examinations. As incentive to participate in the study, students could opt out of two of six course assignments, in exchange for which they would take four, ten-item, multiple-choice quizzes and also fill out a questionnaire following each quiz. The questions on the quizzes corresponded to course content. To motivate them to take the quizzes seriously, the quiz scores counted toward the final semester grade. In total, twenty-one students returned acceptable questionnaires on at least one of the four quizzes, for a total of seventy-six valid and completed questionnaires. Two students' questionnaires were eliminated because of integrity concerns, and one student completed the questionnaires

without completing the pre-requisite quizzes. Six subjects were female, and there was a very diverse representation of ethnic backgrounds as is typical of the student population at this university.

1. This question was difficult to answer.
 - a. Strongly Agree
 - b. Agree
 - c. Neither agree nor disagree
 - d. Disagree
 - e. Strongly Disagree
2. How clear was the wording of this question?
 - a. Very clear—I understood it well on the first reading
 - b. Fairly clear—I had to read it two or three times, but understood it well
 - c. Somewhat unclear—I read it several times, but still I was not confident I understood
 - d. Very unclear—I read it many times but still did not understand it at all
 - e. I did not read this question
3. How many words in this question and answer choices did you not know or understand?
 - a. Zero—I understood all of the words
 - b. One or two
 - c. Three or more
4. If you answered “b” or “c” in the previous question, which words gave you trouble?
5. This was a fair question.
 - a. Strongly Disagree
 - b. Disagree
 - c. Neither Agree nor Disagree
 - d. Agree
 - e. Strongly Agree
6. The correct answer to this question was “C.” If you missed it, please explain why.
 - a. The question was too hard
 - b. The question was confusing
 - c. My interpretation was different than the instructor’s
 - d. I didn’t study/prepare enough beforehand
 - e. Careless error
 - f. Other (please specify) _____
7. Do you have any thoughts, comments or suggestions to share about this question? (open-ended)

Table 3.1—The QuizViz Questionnaire

The quizzes were delivered online via WebCT™, a commercial course-management software package. Using the software, students were allowed access to the

quizzes at specified times, and were restricted to a thirty-minute time limit for completing the quiz once begun. Upon quiz completion, the system provided students with a link to the online questionnaire website. The questionnaire asked seven questions about each of the ten quiz items (see Table 3.1), followed by four questions about the quiz as a whole. The questions asked the students to rate the difficulty of each question, to identify the reasons why they missed questions, and to rate the clarity and fairness of each question. Summary questions asked how long subjects studied for the quiz, how carefully they read the instructions, how clear the instructions were, and what could have been done by either the student or instructor to make the quiz better.

The instructor for this course was a veteran professor with several decades of teaching experience, in particular, in the subject matter of the course. At the time the quiz questions were generated, there were not explicit educational objectives specified for this course. The instructor generated questions by surveying the readings and lecture notes, selecting representative knowledge deemed to be important, and then writing questions thought to assess mastery of this knowledge. The collective experience of the researchers suggests that this is how university professors generally create exams.

3.2.2 Results

Table 3.2 shows the breakdown of questions as they fell into the various Bloom categories. The questions were initially categorized separately by two of the researchers. Disputes over the proper categorization of questions were resolved by discussions with the instructor. Several lessons were learned during the categorization process:

1. It is impossible for an outsider to categorize some questions without knowing the question context. An outsider is not generally able to make the distinction

between Remember and Understand questions which often relies upon whether or not students have been explicitly or implicitly exposed to the material in the question before.

2. Classification was difficult when answer choices assessed different cognitive processes. In at least one case this ambiguity appeared to confuse students resulting in a low number of correct responses.
3. It is likely that some cells in the table will never have any entries. For example, the authors of the taxonomy indicate that Apply pertains only to procedural knowledge, although intuitively an objective asking a student to Apply Conceptual Knowledge, or even Apply Factual Knowledge, seems plausible.
4. In general, the robustness of the taxonomy as a categorization scheme seems an open question. It was a challenge at times to work backward from a question to fit it into a category. In fact, the authors of the taxonomy don't advocate this practice and intend that the taxonomy be used to develop objectives prior to item generation. If done in this direction it seems likely that questions would always fit cleanly into a single category.

Table 3.2 indicates that all of the questions fell into the first two cognitive categories: Remember and Understand. No questions tested cognitive skills at higher levels. Although it was expected that no questions would fall into the Create category (by definition one cannot assess creativity with a multiple-choice assessment item), it was somewhat surprising that no questions were categorized as Apply, Analyze, or Evaluate since there was sufficient course material which would require students to Apply

procedures, or to Analyze or Evaluate interfaces. Several explanations for this lack of variety seem plausible:

1. Multiple-choice questions may not lend themselves well to assessing higher-level cognitive mastery. Although Anderson and Krathwohl's book indicates item formats for the higher-level processes, there doesn't seem to be satisfactory empirical evidence indicating their validity (see Chapter 17, p.298).
2. Without explicit practice or guidance in generating items to assess a variety of cognitive skills, it may be that instructors naturally fall into a habit of writing items that fit a limited number of situations. It is not that items to test higher levels aren't possible, it is just that instructors may not be aware of the differences and hence don't vary their items.
3. The instructor subconsciously wrote items on which it was felt that the students would do well, and hence didn't write questions to assess higher-level skills. This explanation is rather difficult to test, and not supported by the levels of success of the students on the quizzes (the overall average was just over 60%).
4. Questions at higher levels of cognition were not appropriate for the given course content. This does not seem likely, but given that the instructor had never defined explicit learning objectives, it may turn out to be the case. In addition, the instructor did use essay questions in face-to-face administered exams that addressed these higher levels.

The Knowledge Dimension	The Cognitive Process Dimension					
	1. Remember	2. Understand	3. Apply	4. Analyze	5. Evaluate	6. Create
A. Factual Knowledge	2	0	0	0	0	0
B. Conceptual Knowledge	7	21	0	0	0	0
C. Procedural Knowledge	6	4	0	0	0	0
D. Meta-Cognitive Knowledge	0	0	0	0	0	0

Table 3.2—Frequencies of Quiz Questions in Bloom’s Taxonomy Categories (N=40)

Regardless of the reasons for the lack of variety in question types, and the difficulties in categorization, it is believed that a software system that guides the instructor in the development of objectives will result in questions that assess a greater variety of cognitive skills, and which are more accurately categorized.

Question Title	N	% Correct Of:			Discrimination	Score	
		Whole Group	Upper 25%	Lower 25%		Mean	SD
Heuristic vs. Cognitive Evaluations	20	60	85	50	0.31	60.00%	50
User-centered design lifecycle	20	60	71	37	0.37	60.00%	50
User Walkthrough	20	65	100	37	0.61	65.00%	49
Conceptual Model (subway system)	20	65	85	50	0.27	65.00%	49
Paper-clip Icon Example	20	30	42	25	0.38	30.00%	47
Table of Contents	20	90	100	75	0.54	90.00%	31
Affordance definition	20	70	85	50	0.22	70.00%	47
Questionnaire Use	20	60	57	50	0.15	60.00%	50
Cell phone design goals	20	25	57	0	0.59	25.00%	44
Cognitive Walkthroughs	20	70	100	50	0.46	70.00%	47
Overall Mean:						59.50%	

Table 3.3—Results from Quiz #1 as Reported by WebCT™

Another part of the analysis was guided by the students' performance on the quizzes. WebCT™ can produce reports, such as Table 3.3. Items where the percentage of the whole group getting the right answer was less than 45% or higher than 85% were determined to be either too hard or too easy, respectively, and were subjected to closer analysis.

- Which of the following is a valid usability goal for the design of cell phones?
- A. On average, the time it will take a user of the new interface to obtain a voicemail message will be 30 seconds
 - B. Users will find it easy to learn its basic functions
 - C. Users will be completely satisfied with the interface
 - D. 80% of users will find it easy to learn

Figure 3.1—Bad Question Resulting from Assessment of Multiple Concepts

- Of the following usability goals, which of these represents good industry practice for setting such goals for usability testing?
- A. On average, the time it will take a user of the new interface to obtain a voicemail message will be 30 seconds
 - B. Users will find it easy to learn its basic functions
 - C. Users will be completely satisfied with the interface
 - D. 80% of users will find it easy to learn

Figure 3.2—A rewrite of the previous question to correct ambiguity problems

Figure 3.1 gives an example of a bad question, and Figure 3.2 gives one possible rewrite. Only 25% of the students answered this question correctly. During discussion the instructor could not find anything wrong with the question. The answer came when reading the student feedback. One student said, “Our week 3 conference posting asked us to define usability goals. [Our professor] clarified for the class about usability goals, using my posting as an example. The reply was: ‘A key aspect of a usability goal is that it must be measurable.’” Other students made comments such as, “I used personal experience, did not feel A was correct,” “it was an hard question with lots of answer” [sic], and “The wording of each answer was poor.” These comments were the clues that led the instructor to decide that the problem was that there are *two* concepts being tested

here: the concept of *measurability*, and the concept of *usability*. A “valid usability goal” is measurable; this is the concept the instructor wished to assess. However, students only focused on the phrase “valid usability” which made the most common answer B. A guideline which would have helped avoid this problem is to make the answer choices as parallel as possible, only manipulating elements directly related to the concept being assessed. In this case, rephrasing the stem to refocus the attention on the concept of *measurability* or making all of the choices include the phrase “easy to learn” may have helped.


<p>The following paper clip icon that is used to summon help in Microsoft Word is an example of:</p>  <p>A. affordance B. mapping C. conceptual model D. feedback E. is expressed by both items A and B above</p>	<p>A story board:</p> <p>A. is a rough method for describing and laying out the design of a user interface B. is a prototyping method that describes the linkages between the various functions that are being built in the user interface C. is a way of providing help in a user interface that is both painless and fun D. is a method for designing user interfaces by building a story that explains to the user each and every function of the application E. is a rough method for evaluating the early design of a user interface</p>
---	---

Figure 3.3—Bad Questions Resulting from Poor Exemplification of Concepts

Figure 3.3 shows two questions that suffered from the problem of having poor examples to illustrate a concept. In the first question the student must choose the correct concept that applies to the given example; in the second question the student must choose the correct example that corresponds to the given concept. In the first question, the example is not really a good example of any of the concepts, and in the second, several of the choices are correct, requiring the student to ascertain which is the most correct.

Unfortunately, the choices are written too ambiguously for the student to make this assessment.

Two more general conclusions were one, that the difficulty of all of the questions was more or less appropriate, and two, that findings confirm the wisdom of Haladyna's guidelines.

3.3 The Second Pilot Study

3.3.1 Description of the Study

The goal of the second pilot study was to test the procedures that would later be used to evaluate QuesGen. Although it was known that this pilot would not generate enough data to evaluate the chosen statistical procedures, insight into the sufficiency of the other evaluation procedures was desired. If problems were discovered at this phase it would be possible to smooth out these bumps in the road before the full scale test of QuesGen.

The study was run during the final exam period of the Spring 2006 semester at the same medium-sized, urban, northeastern, technical university. Two instructors from the experimenter's department volunteered to try out the QuesGen system. Both instructors were sent an email giving them step-by-step instructions on what to do to participate. Both of the instructors were asked to generate between 5 and 10 MCQs before logging into QuesGen. Then they were asked to download and install the latest version of the Flash player so that they would be able to watch the online tutorial videos. One instructor had a problem with this step that was resolved over the phone. Next the instructors were asked to watch the five tutorial videos. Altogether the length of the videos was 68 minutes. Next, the teachers were asked to write at least five MCQs using

the system. If there was no educational objective, they were asked to write an appropriate objective before writing the question. Unfortunately at the time of the pilot, the student interface to QuesGen had not yet been completed and so the instructors had to deliver the questions they wrote with the system by other means. They had two options to do this: either print the questions out and deliver them on paper, or export the questions in WebCT format, import them into WebCT and deliver them online. Fortunately, one of the professors chose each of the methods so both were tested.

Since the student questionnaires had already been piloted in the first pilot, they were not distributed to students this time. Given the tight time constraints of this study, it was decided that the item analysis reports would be abbreviated. While not ideal, if meaningful results were possible with the abbreviated results, then this would be a good sign that the eventual system evaluation would yield enough evaluative information to be useful as well.

After the exams had been delivered, the exam results were collected from the teachers and entered into the system so that the item analysis reports could be produced.

3.3.2 Results

The two teachers who participated in this pilot used QuesGen to generate five questions each. One of the teachers then delivered the questions to his ten students on paper as part of the final exam. He recounted that he delivered the questions generated by QuesGen on a separate sheet of paper from the other 40 questions given on the exam. The other teacher delivered the questions to his five students using WebCT as a review quiz; since the review quiz was optional, only three out of the five students took the opportunity.

The teachers reported several problems and shortcomings of the system. One teacher experienced a problem upgrading his version of the Flash player. This made it impossible at first for him to watch the tutorial videos. Fortunately he called an experimenter and this difficulty was resolved over the phone. It was decided to recreate the tutorial videos in a version of flash that was less likely to require upgrading so that problems with the version of Flash would not occur in the future. A problem with the automatic upgrade script was also discovered that needed to be fixed.

A second shortcoming that was mentioned was the inability to control the visual formatting of the questions. The teachers wanted the ability to be able to use italics, bold, and other formatting techniques in the text of their questions. While the capability to format the questions by typing in HTML tags existed, it was realized that QuesGen had not included this information in the tutorial videos. Also, even if it had been communicated, asking teachers to learn and remember how to use HTML as part of their questions is not a very reasonable expectation. Actually, while under development, a WYSIWYG (what-you-see-is-what-you-get) HTML editor was included in the system. However, its implementation was buggy and due to time constraints the choice was made to leave it out of the first version of the system. As a result of the pilot, the WYSIWYG tool will be added back to the system, and effort was made to test and debug it thoroughly before the final evaluation of QuesGen.

A third shortcoming was in the documentation of how to export questions for WebCT. The teacher that delivered his questions using WebCT could not figure out how to import the questions into WebCT once they had been exported from QuesGen. Rather than request assistance he decided just to transfer the questions manually into WebCT.

Since the WebCT export function was one of the last to be developed, the documentation for that function had not been implemented yet. In the future that problem is not expected.

One of the teachers volunteered written feedback on QuesGen. Here is one of his comments:

The power of QuesGen is certainly in the question templates, and that is where there is some need for improvement in the next version. The tutorial only discusses Exemplification and Classification. I would like to have invoked additional question templates, but was not quite sure of their unique functionality. Nor do I like going thru them one by one, once I have a question/objective in mind, in order to find the appropriate template which matches the question in mind.

These were very encouraging and useful comments. Indeed, with twenty different question templates to choose from, the task of reading through them and learning how to apply all of them is not small. The teacher suggested a printable template reference manual, preferably with examples of each template being used in a question. This was considered a good suggestion and an effort was made to make the printable manual and examples available for the next test of QuesGen. In addition, the system designers did more brainstorming on how to make the templates more accessible without the need to print anything out.

Another comment by this teacher was very encouraging:

I am very impressed with the capabilities and potential of QuesGen, and have learned a lot about the optimal design of multiple choice questions which I had not known (I have used a lot of "all of the above" and "none of the above" in prior questions!).

This was direct evidence that QuesGen was accomplishing one of its major goals changing the knowledge, beliefs, and attitudes held by teachers about the development of high quality multiple-choice questions. The other teacher involved in the pilot also made

a useful comment in an informal conversation had after he had watched the tutorial. He related that he hadn't realized how bad his old questions were until after he watched the tutorial and started going back over them. These are precisely the same experiences had by the experimenter had when he ran the QuizViz study and began examining his own questions, and they are precisely the experiences QuesGen was designed to evoke.

Other valuable lessons were learned from this pilot. First, the teachers involved did not have a sense, because it was not given to them, of how much time would be needed to interact with QuesGen and complete the tasks that had been set for them. Because of this, it was not possible to collect as much data from them as originally planned, the reason being that the teachers had not left enough room in their schedules to accommodate all of the experimenter's requests. First of all, the invitation to participate in the study did not mention that watching tutorial videos was involved; hence, they did not factor in the extra hour needed to complete this task. Also, since the completing of the question quality checklist was not a part of the process that they could have foreseen, one teacher did not use the checklist at all, and the other teacher used it for only two out of the five questions that he wrote. This realization prompted a change in the orientation to the experiment planned for the full QuesGen evaluation. The full evaluation included a list of all the tasks that would be requested and an approximate amount of time that it would take each teacher to complete that task. Despite their time constraints, however, the teachers did take the time to generate educational objectives with which to associate their questions.

CHAPTER 4

AN INFORMATION SYSTEM MODEL FOR TEACHER CHANGE

4.1 Introduction

The first stated goal of the literature review was to use existing published research to help define the requirements of an information system that might help teachers to become better at assessment. Having completed that literature review, this section will now synthesize the literature into a coherent information system model. First, an important caveat, it is far beyond the scope of this thesis to completely implement and test this model. QuesGen, the system that was built for the thesis, will implement only a subset of the functionality indicated in the model. If QuesGen is successful at achieving its goals, the plan for the future is to evolve QuesGen to the point that it will allow an evaluation of the full model that is presented in this chapter.

This chapter is organized as follows. First, Section 4.2 describes the five atomic elements of the model. Section 4.2.1 describes the core of the system, which is based upon Fishman et al.'s KBA framework for teacher professional development. KBA's core claim is that changing teachers' knowledge, beliefs and attitudes will bring about positive teacher change. Section 4.2.2 describes how the stages, processes, and metrics of change defined in the transtheoretical model of behavioral change can be used to bring about changes in KBA in an intentional and measurable way. Up to this point, the system model describes a very general model for teacher change, but does not specify any specific changes. Section 4.2.3 will discuss how to tailor a software system to focus on a specific set of knowledge, beliefs, and attitudes by taking advantage of the established

research in a specific sub-domain of education. As example the use of Bloom's Taxonomy and Haladyna's work on MCQs as the sub-domain of choice for the system will be discussed. Now, with a thorough understanding of the target of change and the strategy for change, Section 4.2.4 will bring to bear the specific, relevant expertise from the Information Systems realm to inform how to convert the ideas in the previous sections into digital artifacts that can make up a coherent software system. The last element of the model addresses the need for such systems to interact with the outside world—Section 4.2.5 describes the use of the IMS Global Consortium's data models to provide interoperability with other learning management systems (LMSs) on the market. Finally, Section 4.3 synthesizes all of the above elements into the complete model. The chapter following this one will describe how the elements of this software model were expressed in the building of QuesGen.

4.2 Elements of the Model

A simplified representation of the model which contains all of the major elements is shown in Figure 4.1. The five elements, the first four of which come directly from the

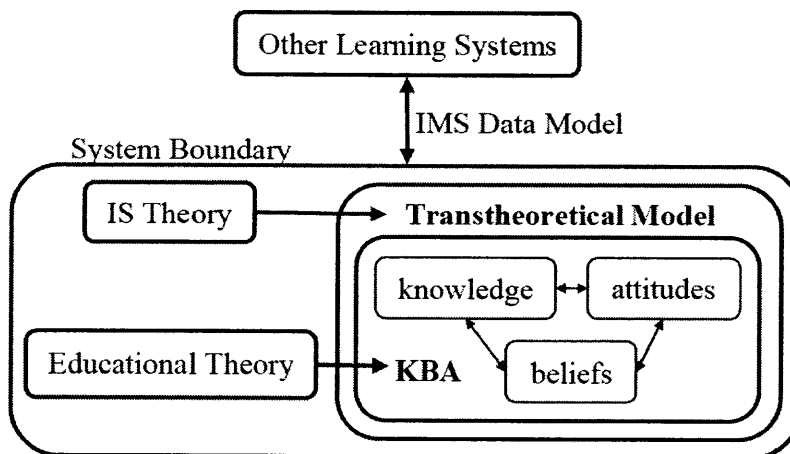


Figure 4.1—Elements of the system model

literature review, are: KBA, the Transtheoretical Model (TTM), Educational Theory, IS Theory, and the IMS Data Model.

A moment will be taken to distinguish the model from a “software architecture” as defined, for example, in Fielding (2000). Software architectures are limited in scope to defining components, all of which can be represented as code. A quick inspection of the model suggests that this is not the case. The model describes the issues that a system designer would need to address to build a working system that addressed some aspect of teachers’ professional development. This model gives system designers guidance as to where and how to incorporate the domain expertise of the various people who may be involved in the building such systems, some of whom will be software architects and will develop an underlying architecture for the system. Using this model the specification of concrete functionality that would comprise the resulting software should be relatively straightforward, though non-trivial.

Figure 4.2 shows the roles that would need to be filled in a team assigned the task of designing a professional development (PD) system based on this model. These roles may be filled by four separate people, or some people may play more than one role. In the case of QuesGen, all four roles were played mostly by a single person. The *education specialist* would provide expert guidance in what research has proven to be the

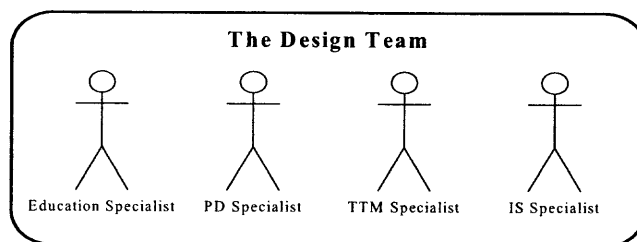


Figure 4.2—A System Design Team for a PD System

most appropriate knowledge, beliefs, and attitudes (KBA) in the target behaviors that the system seeks to change. The *PD specialist* would devise a plan to incorporate that KBA into teachers' everyday practice. The *TTM specialist* would be responsible for designing appropriate behavioral interventions composed of the various processes of change that TTM describes to accomplish the goals set out by the PD specialist. The TTM specialist would also devise metrics to discover and describe teachers' state of change. The *IS specialist* would then bring the research in Information Systems to bear in designing software artifacts that will implement the interventions specified by the TTM specialist. The resulting specification could then be implemented as software.

The resulting system may be a standalone system, or it may become a component of a larger system which handles other aspects of the learning process. Ultimately the goal would be to develop a full complement of tools built according to this model that would cover most of the activities required of teachers on a day to day basis. Teachers using this system would gradually improve as the various aspects of their practice came into line with the "best practices" of the profession as discovered by education research. Such a system will likely take many years to develop. This dissertation is the beginning of that effort.

4.2.1 The Core—KBA

The core realization guiding the model is that the most reliable way to improve teaching is to focus on the knowledge, beliefs, and attitudes that underlie teaching practice. As described in the literature review, this realization was reached through studying years of successful and unsuccessful teacher professional development programs, and pulling out the common thread that was woven through all of them (Fishman, et al., 2000). Black et

al. (2003, p91) echo this idea, and Fishman et al.'s KBA framework received some validation through recent case studies (Kubitskey and Fishman, 2005). The idea is also compatible with Shulman's (1986) influential research into the impact of teachers' pedagogical and content knowledge on their effectiveness. It is useful, therefore to revisit Kubitskey and Fishman's description of effective professional development:

[Effective professional development programs (PD) are] well planned over the long term using evaluations of past PD opportunities to inform future PD. The PD is structured around a specific need of participants, proximal to teacher practice and supplying usable information to teachers. The participants should either already be in a community or should have common ground for forming a community, such as teaching the same unit, teaching in the same school, etc. The PD should take place over an extended period of time, to promote the creation of a community of practice whose participants have common goals addressed by the PD. Finally, the PD activities themselves must offer opportunities for teachers to engage in inquiry of either content or pedagogy. (Kubitskey and Fishman, 2005, p2)

The reference above to meeting teachers' needs provides the insight for a key assertion of this system model:

Assertion 1: A system designed to help teachers change their behavior should always implement as its core functionality some tool or set of tools that help teachers accomplish a task or tasks central to their role as teachers.

One of the reasons that Kubitskey and Fishman wrote the above statement is that a characteristic of unsuccessful PD has been found to be programs that seem irrelevant to teachers, and which don't immediately address their needs or concerns. One of the most well-known findings in software use is that if users do not perceive that a system will be useful to them, they are significantly less likely to use it (Venkatesh, et al., 2003). Because of these findings, Assertion 1 means that not only should the core functionality of a PD system support central teaching tasks, but also that the usefulness of this tool

should be immediately and apparently obvious to teachers. Failure to make a tool's usefulness obvious will likely result in lower rates of adoption.

A second assertion follows from another of Kubitskey and Fishman's statements, those related to the duration of successful PD programs:

Assertion 2: Systems designed to help teachers accomplish tasks central to their role as teachers, should be designed as systems that will be used on a regular basis.

If a software system is to be used as the basis of a PD program that takes place over an extended period of time, it makes sense that the system should be designed with regular use in mind. The design parameters of such systems are different from systems that are designed to be used only infrequently. This will be made clear with an example.

Take, for example, tax preparation software. Most people only fill out tax forms once a year, which is not often enough for them to learn how to accomplish this task without making explicit reference to the instruction manual. Tax preparation software, therefore is designed as a "wizard" that walks the user step-by-step through the process, asking them key questions at each point and providing additional clarification whenever a question is ambiguous. Tax professionals, on the other hand, are more likely to use software that allows them to enter numbers into tax forms directly, without unnecessary interviewing and explanation. This extra "hand holding" functionality is more likely to just get in their way. Software designed for a person who is not yet, but on their way to becoming, a tax professional would have some combination of the two types of software, i.e. wizards for the novice that can be hidden or turned off as the novice becomes a professional. Following this logic, systems designed for teacher PD must conform to the requirements of this third type of system. In other words, the expertise required to complete the task associated with the system must be built into the system in a way that is

obvious and easily accessible, but which can be turned off or hidden when it is no longer necessary.

To summarize this section, the core of the model sets up several key requirements.

PD systems based on the model should:

- Be based on inculcating a core set of knowledge, beliefs, and attitudes,
- Implement functionality central to teachers' everyday tasks,
- Implement functionality whose use and purpose is obvious to teachers, and
- Provide scaffolding for teacher learning.

Other implications of Kubitskey and Fishman's statements, particularly those related to community, will be dealt with in other elements of the model. These requirements partly establish what the system should do. The next element in the model, the organizing framework, will explain how to realize these requirements.

4.2.2 The Organizing Framework—The Transtheoretical Model

A problem with the KBA framework is that it does not specify a concrete set of mechanisms by which PD can be accomplished. Since the goal of PD is ultimately to change people's behavior, transtheoretical model of behavioral change (Prochaska and Velicer, 1997) has been adopted as the organizing framework for the model. As discussed in the literature review, the merits of TTM are that it:

- Provides a comprehensive and concrete set of interventions designed to bring about behavioral change
- Is general enough to be adapted for most types of behaviors both at the individual and organizational level (Prochaska, et al., 2001a), and

- Specifies not only the mechanisms of change, but also the means to measure and monitor change and make reasoned decisions about which specific interventions to employ at a given stage of change.

Before discussing the implications of TTM for the model, the three major elements of TTM will be briefly described again—the stages of change, the processes of change, and the metrics of change—and also how each of these elements is realized within the model.

4.2.2.1 Stages of Change

TTM describes six stages of change, which were discussed earlier in the literature review (see Section 2.4.5): precontemplation, contemplation, preparation, action, maintenance, and acquisition. One of the practical aspects of any PD program is that for any target population, some of the teachers will already know about and have adopted the particular behavior that is being espoused by the PD program. On the other hand, there will be teachers to whom the practice has never occurred, and yet others who have heard of the practice, but are opposed to it for some reason. Still others might know about, or be in the process of changing their teaching to conform to the practice, and they may be struggling with it. In terms of TTM, each of these teachers is in a different state of change. TTM holds that the behavioral intervention strategy that works for all of these teachers will be different because research has shown that different processes of change are more or less effective at different stages in the change process.

A PD system designed with the model would envision the characteristics of teachers at various stages of change and modify itself in accordance with those stages. Part of the design process would be to do a scenario-based design using profiles of

teachers in various states of change (Carroll, 1999). A significant design challenge will be to encourage teachers to use the system who are initially hostile to the KBA that it manifests. Given those stages, then, the system would bring to the fore the functionality that is designed to implement the mechanisms described in the ten processes of change.

4.2.2.2 Processes of Change

The ten processes of change, which are defined earlier in the literature review (see Section 2.4.5), are: consciousness raising, dramatic relief, self-reevaluation, environmental reevaluation, self-liberation, social liberation, counterconditioning, stimulus control, contingency management, and helping relationships. Each of these processes is associated with concrete interventions that can be used to help people change their behaviors. For example, consciousness raising is defined as informing the teacher of concrete facts and figures related to the behavior in question so that they understand the pros and cons associated with changing or not changing their behavior. It is relatively straightforward to envision functionality in a system that provides this type of information, such as help screens, and tutorials.

Once the target KBAs have been identified, and once scenarios for teachers in all of the various stages of change have been developed, it is possible for the system designers to go through the list of processes and then design software functionality that will implement those interventions. Kubitskey and Fishman wrote about the importance of forming a community around the given behavior to be developed as part of the PD. In the context of TTM's processes of change, these communities would fall into the category of helping relationships, and in software might be realized in online or virtual communities. Likewise, it will be possible by using the list of interventions to target all

of the interventions that can be implemented as software. It will also help uncover areas where no software solution is possible and some other means of implementation must be used.

Since the processes of change are more or less effective depending on the stage of change of the teacher, the final problem in TTM is how to determine in what stage of change a person presently resides. Fortunately the model addresses this issue.

4.2.2.3 Metrics of Change

TTM defines three metrics that are used to determine in what stage of change a person currently resides: decisional balance, self-efficacy, and situational temptation. The general definitions of these three metrics were discussed in the review of the literature. These three metrics take the form of short questionnaires, and must be defined and validated for each new behavior that one seeks to change.

Looking at examples of concrete instantiations of these measures published on the CPRC website (<http://www.uri.edu/research/cprc/measures.htm>, 5/12/06) one can see that the questions asked in these instruments, and the resulting interpretations, are fairly straightforward. Since the questionnaires are made up primarily of closed-ended, binary choice (i.e. yes/no) items, their implementation in software is also straightforward. The interpretation of the results can also be automated (Prochaska, et al., 2001b). The action taken on the basis of these interpretations would be to reconfigure the functionality currently available in the system, or to suggest to the teacher that he or she learn about new sets of functionality.

Developing these metrics for teachers' stages of change may be one of the most challenging aspects of designing a system based upon the model. However, if this

element of the design is successful, then over time it will generate a very high degree of data and insight into the processes of teacher change.

4.2.2.4 An Implication of TTM

A problem with many PD programs is that they take a “just add water” approach to developing teachers. These programs deliver pre-packaged resources that are “ready to use.” This sounds nice to teachers because ostensibly it means that they will have less work to do to implement these new ideas. However, what such programs don’t take into account is that in delivering a finished product, they short circuit the growth teachers experience by wrestling with an educational problem. A good example of this is McKeown and Beck’s (2004) description of their effort to develop some PD resources called Accessibles. Accessibles were brief, single-issue documents which teachers could use to implement a teaching strategy called “questioning the author.” The problem with this study is that it was clear from the description of the study that the real benefit to be gained from the Accessibles was not in using the Accessibles created by someone else, but in actually creating them oneself. The act of creating the Accessibles forced a teacher to really wrestle with the issue of what made the practice of “questioning the author” viable as a teaching strategy. This struggle with new pedagogical ideas is the kind of experience that leads to a change in knowledge, beliefs, and attitudes. However, many PD programs undercut this by providing pre-packaged teaching resources in an attempt to be helpful.

Restated, the criticism is that since the recipients of these resources aren’t involved in their development they may not understand or be persuaded to accept their deeper significance or value, and hence not employ them. This view is supported by the

findings of Kubitskey and Fishman's (Kubitskey and Fishman, 2005) case study. They taught a new concept to 28 teachers, and observed their beliefs, knowledge, and practices both before and after the training. In analyzing the resulting patterns of change in knowledge, beliefs, and practices they reached the conclusion that:

*If the professional development convinces teachers of the value of adopting what the PD is teaching, teachers make the change in practice.
(p25)*

Systems built using the model will take this basic premise into account and avoid being simple repositories for pre-packaged content.

The transtheoretical model approaches each person as a rational individual. Systems built using TTM as the organizing framework have a high degree of transparency since the rationale for any behavior advocated by the system is explicit in the functionality. Teachers using the system are not asked to blindly implement a technique because "it works." Rather, an effort is made to convince and teach a person that the new behavior is not only worth having, but demonstrably better than rival behaviors. It should be possible for a teacher using the system to find the research which supports each function. As a result, teachers using systems developed with this model should actually become better teachers because they will have deeper understanding of both pedagogy and content.

4.2.3 The Education Domain

The role of educational research would be central in the selection of the specific Knowledge, Beliefs, and Attitudes that would make up the core of the system. For example, as part of the research for this thesis a system was designed and built to help teachers write better multiple-choice questions. As demonstrated in the literature review,

this choice for the construction of a system is warranted based upon research into the importance of assessment in the learning process, and also research showing the dearth of competence that teachers possess in the area of assessment, particularly in writing MCQs.

As Leidner and Jarvenpaa (1995) indicated in their review, much software for education merely automates education-related tasks without considering the opportunity to take advantage of software to innovate and improve upon these processes. In answer to Borko's (2004) call to have a higher degree of basic research incorporated into teacher professional development, this software model would require the use of educational research to support the inclusion of any core functionality. Conversely, it is a goal that the motivation to design new tools in the first place come from research-based insight into how people learn, and the best methods for teaching.

As an example of the incorporation of research from the education domain into system design, the implications of Haladyna's work on effective MCQs is briefly described (Haladyna and Downing, 1989a; Haladyna and Downing, 1989b). These implications are more fully explained in the next chapter where it is described how this research was realized in the design of QuesGen. Haladyna's research has focused on empirical validation of the features of MCQs that are most likely to make a question successful. The first basic idea is that the goal of any successful MCQ is to learn as much as possible about the current state of a student's knowledge. The second basic idea is that whenever a student responds to a question, there is always a chance that the student will get the answer correct or incorrect *NOT* because of his or her understanding of the subject, but because of some flaw in how the question was asked.

Unsuccessful questions cause the people interpreting them, i.e. students, teachers, parents, administrators, etc., to draw incorrect conclusions about the degree to which learning has occurred. The seriousness of the consequences of misinterpretation depends upon the gravity of the test. For a high-stakes test, misinterpretation can lead to a failure to gain admission to college, a failure to graduate from high school, or can cause a school under the No Child Left Behind Act to be labeled as “failing” and therefore ineligible for federal funding. From these examples, it is easy to understand why much research has focused on understanding how MCQs perform. But even in a low-stakes environment such as classroom teaching, if a teacher incorrectly concludes that students have mastered a given concept and goes on to the next one, the students in need of remediation may become frustrated or confused (see the end of Section 2.5.1.2 for an anecdote that illustrates this concept). This makes effective learning difficult, at best. As such, it is important that teachers who choose to use MCQs for assessment make an effort to use effective MCQs.

Haladyna has shown that MCQs are complex to construct because there are *many* variables that determine the quality of a question, such as its length, the vocabulary used, the layout on the paper or screen, the correctness of the grammar, the salience of the topic, and the novelty of examples used, just to name a few. His research summarizes the empirical research on all of the factors known to influence question quality. As it was learned from the literature, teachers are highly unlikely to be familiar with these findings. Also, as it was discovered in the review of the currently existing software for developing MCQs, no heed is paid to Haladyna’s findings in software tools. The tools available do not offer any guidance to the teacher in how to make important decisions like how many

answer choices to develop, or whether or not the use of ‘all-of-the-above’ is advised. On the other hand, these tools do provide support for the development of questions with features specifically contraindicated by Haladyna’s work, such as the use of the multiple-choice format, or questions with answer choices laid out horizontally as opposed to vertically.

Hopefully at this point, the degree to which expert knowledge from the educational domain can inform the design of software to support teaching practice is relatively clear. The specific implementation of Haladyna’s findings into concrete software functionality will not be described here. This description happens in the next chapter which describes QuesGen’s features in detail.

4.2.4 The Information Systems Domain

Next the role of the information systems professional in the design of systems to help teachers develop their professional skills is addressed. The role of the IS professional is to bring relevant research and expertise in the development of software systems to bear upon the problem of how to implement the functionality specified by the other members of the design team. Again Leidner and Jarvenpaa’s warning is relevant—the IS professional should help ensure that the tools developed for teachers innovate, and do not just automate, teaching practice. This is the point in the design when structuration theory and captology (computers as persuasive technology) exert influence over the design. This is also the point in the design where understanding of user interface design becomes crucial to the success or failure of the system to achieve its goals.

For systems built with this model, the goal of the system will always be to change the knowledge, beliefs, and attitudes held by a teacher with the intent of changing the

way that the teacher practices his or her craft. As shown in the research on teachers' professional development, successful change requires not only that teachers be aware of new techniques, skills, or ideas related to teaching and learning (the knowledge), but they must also be convinced that these techniques, skills or ideas are the best ones to implement (the belief), and also that they possess the competence or ability to act on these techniques, skills, or ideas (the attitude). The transtheoretical model serves as a mediator in that it can suggest to the IS professional the mechanisms by which the KBA can be impacted. Given these goals, the role of the IS professional will be illuminated by an example of how captology and structuration play a role in system design using QuesGen as the focus. As in the previous example, the more explicit description of QuesGen's design will be saved for the next chapter and only enough detail to demonstrate this element of the system model will be given here.

DeSanctis and Poole (DeSanctis and Poole, 1994) argued that the structuration model most be looked at from the level of small groups within an organization. In their formulation, which they referred to as adaptive structuration theory (AST), groups adopt technology in ways that conform to the expectations and social norms of the group. In other words, they may not use technology in the way that the developers intended, but rather adapt it to fit their own way of viewing their task. AST would predict, then, that if a software tool is too autocratic in dictating the means by which a task is done, and if that means was not in alignment with the pre-existing norms of a particular group, then the chances that the tool would be successfully adopted are small. However, if the tool is flexible enough for the group to make it fit their particular interaction style, then the chances it will be adopted increase. That is not where the story ends, however. Once

having been adopted, the tool begins to exert an influence on the group to change its members' patterns of interaction. In time, the social norms of the group will change and their use of the tool will also evolve. The challenge of AST is to use the understanding of this role of tools to design them in such a way as to bring about positive change in organizations. Let's apply these ideas briefly to QuesGen.

Teachers and students will serve as the group whose ideas about assessment are the focus of interest. Beginning with teachers, it must be assumed that teachers already have some knowledge, beliefs, and attitudes toward their assessment practice. This understanding governs the way that a teacher will write, distribute and score questions, the way he or she will interpret the results, and also govern the resulting interaction between the teacher and students when presenting and discussing the results. Research indicates that these assessment-related practices will be significantly different than the ones that would be considered "best practice" when held up to empirical standards, such as those developed by Haladyna. If the goal is to change the interaction between teacher and students surrounding the use of MCQs, then the tool needs to help the teacher to understand the means, and believe in the wisdom of following Haladyna's guidelines. One way to do this would be to build the functionality of the tool rigidly, so that the only way to be able to use the system would be to do it the "right" way. AST would predict that if the difference between the "right" way and the teacher's way is too dramatic, then the teacher will not use the tool. Therefore, AST would advocate a softer approach which supports, but does not rigidly enforce, the development of MCQs according to Haladyna's criteria.

Once teachers have been convinced to use the tool, however, the tool can begin to exert an influence back upon the teacher. Through the use of elements in the interface, for example, it may be possible to entice the teacher to explore using some new functionality that promises to make him or her more effective at writing questions. One idea for this, which will be described more thoroughly in the next chapter, is the use of a question quality checklist. Putting the checklist in a prominent position on the page where questions are authored will send a strong message to the teacher saying, “Hey! Here is a list of things that you might want to check off before saving your question.” As each item on the list is examined, understood, and then put into practice, the teacher’s knowledge will be increased. When the results of the quiz come back and the teacher is able to see a difference between the “new” questions and the “old” questions (i.e. those created with and without the checklist), his or her beliefs about the wisdom of using the checklist will be influenced, as well as the attitude toward using the list. In this way, over time, the tool can have an impact on the teacher’s practice. The goal of the tool, then, is to maximize the chance that it will have a positive impact on teaching practice by being suggestive, but not overly coercive. Again, the details of how this interaction takes place are left vague here, as their detailed explanation will occur in the next chapter. It should be clear, though, from this description how this theory leads into the next one that influences the design of the tool.

Captology, the study of computers as persuasive technology, offers guidelines for how to implement a system that is suggestive without being overly coercive. At its simplest, captology suggests that a computer system that is “helpful,” i.e. it fosters feelings of competence and success in the user, will be more likely to be used, and used

for longer periods of time than a similar system that is “unhelpful.” Again, the expression of these theories in QuesGen will be further detailed in the next chapter, but is included here to flesh out the discussion of the elements of the system model.

The IS professional plays a vital role in bringing knowledge of computers and social systems to bear upon the problem of developing teachers’ professional skills. Chiefly, this member of the design team will evaluate the trends in technology and in the literature on implementation to devise an innovative and compelling realization of the principles developed by the education specialist, the PD specialist, and the TTM specialist. One piece of the puzzle remains, and that is the link to the outside world.

4.2.5 The Outside World—Interoperability with IMS Data Models

A key requirement of any modern system is going to be the ability to import and export data flexibly in a standard format that promotes compatibility between systems. This framework advocates the use of the data models developed by the IMS Global Consortium. IMS describes itself on its website as follows:

The mission of the IMS Global Learning Consortium is to support the adoption and use of learning technology worldwide. IMS is a non-profit organization that includes more than 50 Contributing Members and affiliates. These members come from every sector of the global e-learning community. They include hardware and software vendors, educational institutions, publishers, government agencies, systems integrators, multimedia content providers, and other consortia. The Consortium provides a neutral forum in which members with competing business interests and different decision-making criteria collaborate to satisfy real-world requirements for interoperability and re-use.

IMS develops and promotes the adoption of open technical specifications for interoperable learning technology. Several IMS specifications have become worldwide de facto standards for delivering learning products and services. IMS specifications and related publications are made available to the public at no charge.

IMS standards are of high quality and reliance upon them in the building of systems using the model increases the likelihood that they may be later incorporated into one or more of the major LMS projects on the market or otherwise be compatible with them.

4.3 Combining the Elements

As with any model, the goal is to take a large and complex project and organize it in such a way that the construction of the modeled edifice is sound, well-planned, and complete. As indicated earlier, the use of this systems model is designed to be straightforward, though non-trivial. Figure 4.3 shows the expanded system model including all of the stages of change, and processes of change from TTM, and also illustrating the means in which educational and information systems knowledge would be brought to bear in the design of a specific system designed to implement teacher professional development—QuesGen in this case. In the next chapter, QuesGen will be described—a tool to inculcate best practices for writing multiple-choice questions in teachers, that partially implements this systems model.

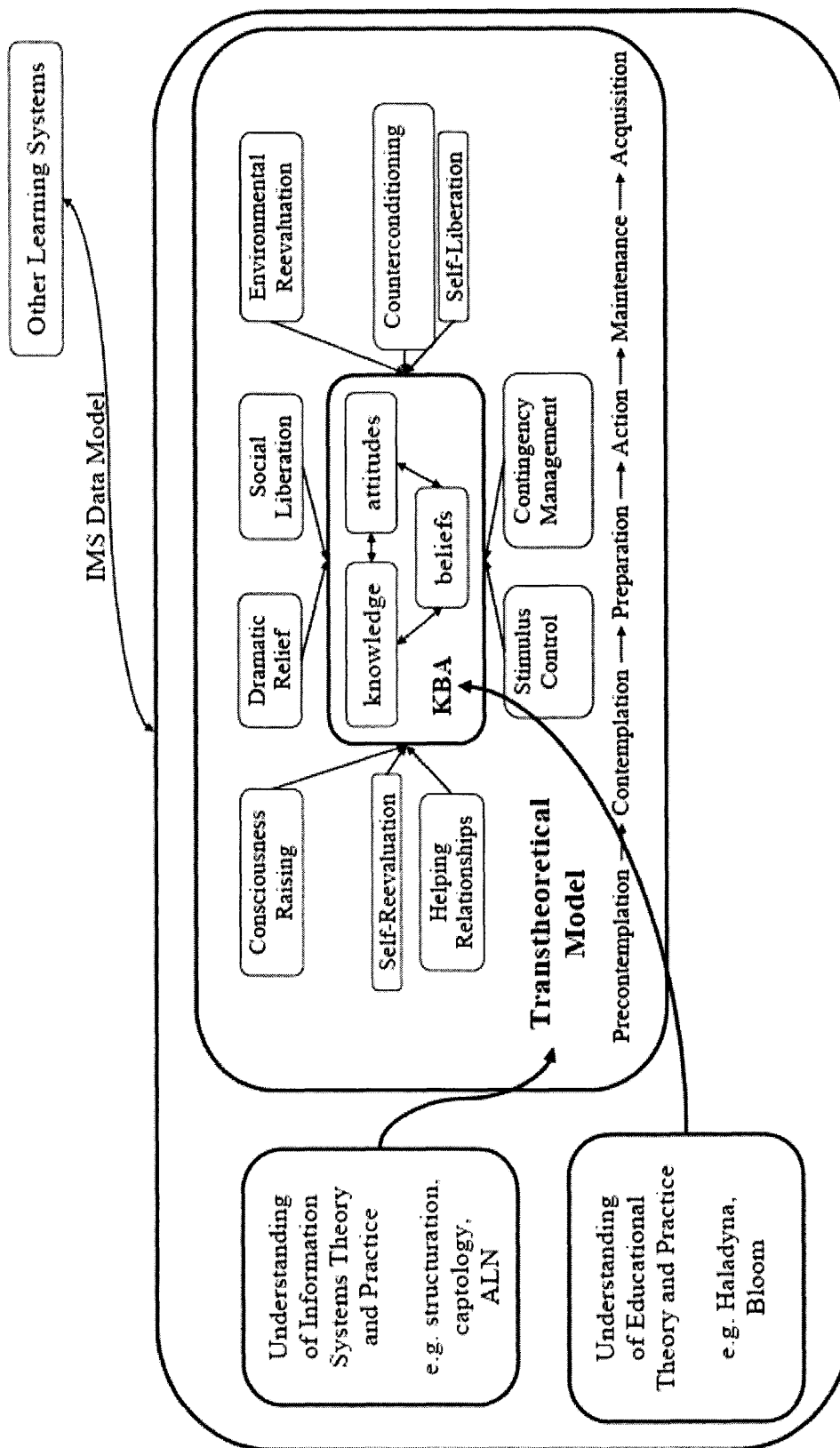


Figure 4.3—The Expanded System Model

CHAPTER 5

DESCRIPTION OF QUESGEN

5.1 Introduction

QuesGen is a web-based software tool designed to help teachers write better multiple-choice questions (MCQs). QuesGen's functionality is designed to inculcate in teachers the proper knowledge, beliefs, and attitudes to foster best practices in the writing and use of MCQs. The scope of functionality provided by QuesGen is likely to be too specialized to allow the system to stand on its own as a separate piece of software. Therefore the QuesGen tool is not currently (but is designed to be) incorporated as a module that would be part of a larger learning management system (LMS) such as WebCT (webct.com), Blackboard (blackboard.com), Moodle (moodle.org), or Sakai (sakaiproject.org). The tool is designed to be one of a group of tools that have the goal of inculcating formative assessment practice in instructors. By itself, it is not expected that the use of QuesGen will cause a radical change in instructors' teaching practices, but it is hoped that awareness of the importance of assessment quality will be increased via interaction with the tutorial elements embedded in the tool. As such, the tool has been designed to guide question developers through a process that represents best practice in question development, and by nature, focuses the attention of instructors upon the core learning goals implicit in a given unit of instruction, and conversely, away from the more trivial elements involved in that unit. This is in keeping with the earlier discussion highlighting the importance of objectives for effective learning.

The process embedded in the tool has been taken from the work of Haladyna (2001), and represents the synthesis of a great deal of research into the development of high quality assessment items. The steps in the process are as follows:

1. Determine the question objective
2. Select an appropriate question template
3. Develop the stem and answer choices
4. Revise it to remove flaws
5. Administer to students
6. Review student performance and feedback
7. Revise, Discard, or Keep the question for further use/modification

Extant question generation tools in the LMSs listed above are much simpler and follow a process which typically involves steps 3 and 5, and sometimes step 6 above.

This chapter will not be a typical software description. QuesGen was designed to implement the elements of the software framework for teacher change, and its functionality will be described in terms of those elements. The processes of change from the transtheoretical model will be used to organize the sections of this description. Since functionality does not exist that addresses all of the processes, this section is generally subdivided into implemented processes and non-implemented processes. Following this is a more traditional description of the system using data models and use cases.

5.2 Implemented Processes of Change

The implemented processes of change were consciousness raising, counterconditioning, self-reevaluation, environmental reevaluation, self-liberation, and social liberation. The following topics are discussed for each process:

- The part(s) of KBA that is(are) targeted (i.e. Knowledge, Beliefs, Attitudes)
- The source of any associated theoretical motivation from education
- The source of any design wisdom from information systems
- The functionality designed to implement that process

Although it has been covered previously, a brief description of the process of change is also included in each section.

5.2.1 Consciousness Raising

As described in the transtheoretical model these interventions are designed to provide concrete facts and figures related to the behavior in question. The individual will learn more about the pros of changing and the cons associated with not changing one's behavior. In terms of the KBA framework the knowledge QuesGen seeks to impart is about how to identify, create, and use high-quality multiple-choice questions, and conversely, the problems associated with using low-quality MCQs. QuesGen also seeks to change typical beliefs about the purpose of MCQs—that they are primarily for learning more about your students than ranking them or giving them a grade. Three features that were implemented with consciousness raising in mind were the online tutorials, online help, and the screen layouts, particularly the question entry form.

5.2.1.1 Tutorials

Tutorials are delivered in one of two formats: audio recorded over PowerPoint slides, or as “screencasts” (videos of action on the computer screen with voice narration). The choice of this medium was motivated by literature from the asynchronous learning networks field (ALN) which shows that this method of receiving instruction online is preferred over pure text or “talking head” videos (citation?). Figure 5.1 shows a screenshot of one of the tutorials incorporated into QuesGen. Tutorials range in length from about eight minutes to about twenty-two minutes, and teachers can use the slider to move to any point in the tutorial at any time. Tutorials were developed using the Flash video format. Flash support is installed in over 97% of web browsers (<http://www.adobe.com/>). The content for the tutorials was based upon the literature on

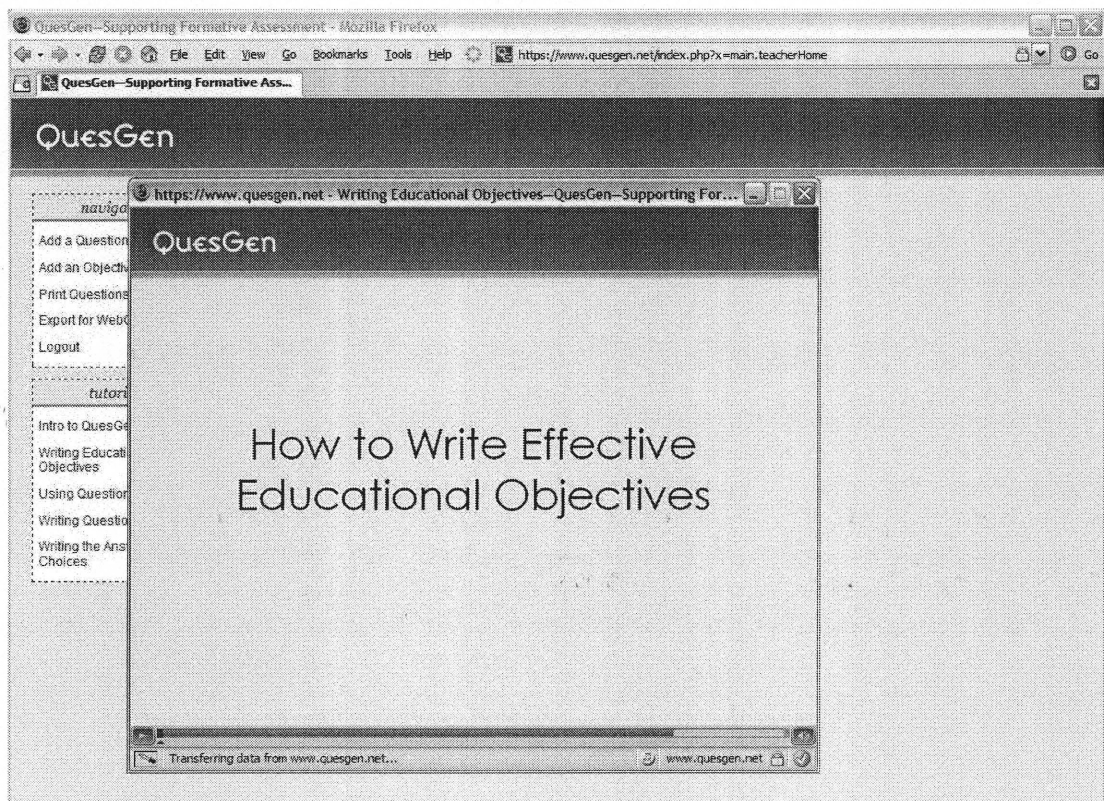


Figure 5.1—A Screenshot of a QuesGen Video Tutorial

educational objectives, question templates, and MCQ development reviewed earlier. Tutorials can be accessed from the teachers' home page, or in context by clicking on the tutorial icon (📖).

5.2.1.2 Online Help

Tutorials were reserved for content where it was deemed that a longer, more involved explanation of a concept was necessary. For situations where a concept or feature could be explained with a smaller amount of text, online help was used. Figure 5.2 shows an example of a popup window containing online help. The majority of the online help was focused on explaining the elements of the question quality checklist. The checklist, which will be described in a moment, was taken from Haladyna's work on features of

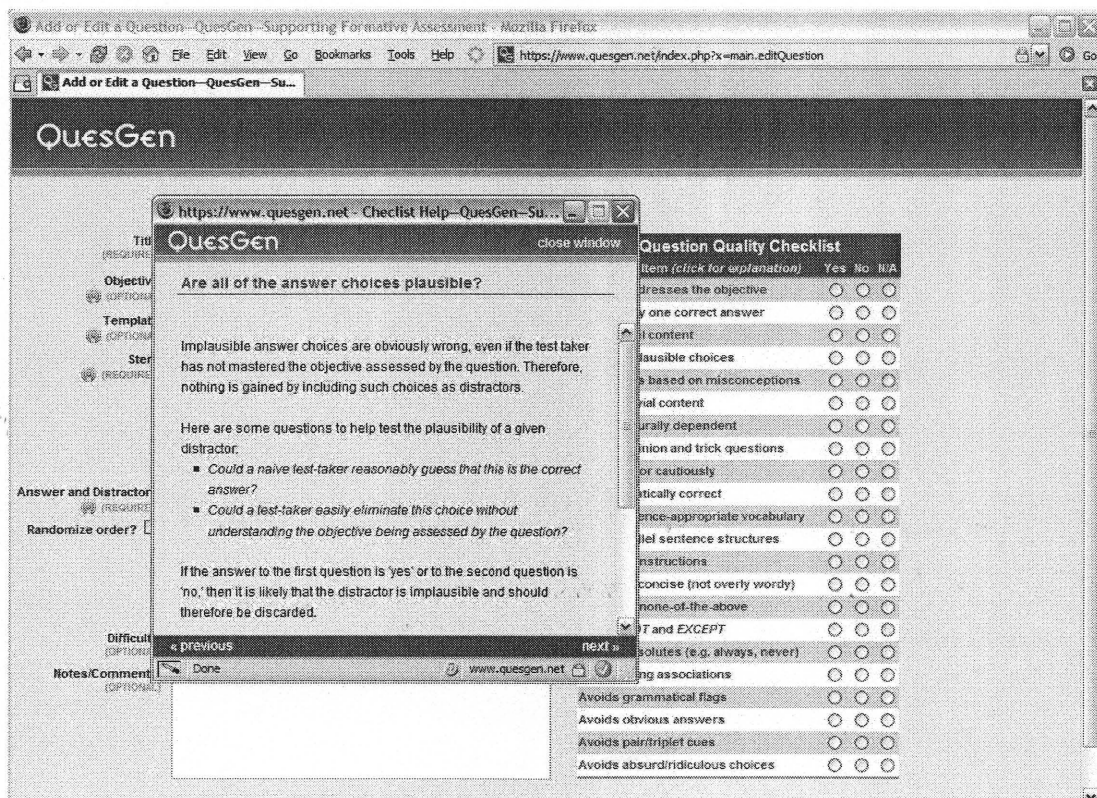


Figure 5.2—A Screenshot of an Online Help Popup Window

effective MCQs. It was not expected that teachers would necessarily understand the meaning of all of the elements of the checklist, and so a link to a help screen for each item was developed. Using the “previous” and “next” links in the popup help screen allows teachers to move through the list of items without having to close the help window.

5.2.1.3 Screen Designs

One of the most important elements of consciousness raising incorporated into QuesGen is the screen design. The screens, particularly the question entry form, were developed so as to demonstrate to the teacher that there were more elements to a successful MCQ than just the stem and answer choices; a successful MCQ is aligned with an explicitly stated educational objective, may follow an established semantic template, and conforms to the best practice guidelines

Question Quality Checklist			
Quality Item (<i>click for explanation</i>)	Yes	No	N/A
Clearly addresses the objective	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has exactly one correct answer	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Uses novel content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has only plausible choices	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distractors based on misconceptions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids trivial content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is not culturally dependent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids opinion and trick questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses humor cautiously	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Is grammatically correct	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses audience	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses parallel	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has clear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing is	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids all	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids M	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids at	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids cl	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids grammatical flags	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids obvious answers	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids pair/triplet cues	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids absurd/ridiculous choices	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

click to drag

Question Preview

O: Applications of do...while loops

Q: In your program you need to send a message to person on the e-mail list. Which one of the following looping methods is most appropriate?

A. do...while

B. for loop

C. while loop

evident in the question quality checklist. The **Figure 5.3—Question Quality Checklist** unsubtle goal of the screen designs was to promote a sense of curiosity and a fuller understanding of the many and sometimes complex facets of a good MCQ. The question quality checklist, shown in Figure 5.3 is a good example of this principle. The checklist was a prominent feature, taking up the entire right half of the screen. The items on the checklist were designed so that, once familiar with them, a teacher could quickly run down the list for each new question and double check to make sure that all of the

guidelines had been followed. As can be seen in the screenshot, this teacher has identified a potential flaw in this question—the “has exactly one correct answer” checklist item is set to “no,” perhaps indicating that, on second thought, the teacher might need to revise the chosen example so that there is only one clearly best answer choice. That each checklist item is a link is supposed to promote exploration and discovery. Clicking on one of the links in the checklist rewards the instructor with potentially helpful information about that particular checklist item.

The screenshot shows the QuesGen web application interface for adding or editing a question. The interface is divided into several sections:

- Add or Edit a Question:** This section contains several input fields:
 - Title:** A text box containing "Identify do...while scenario".
 - Objective:** A dropdown menu with "Applications of do...while loops" selected.
 - Template:** A dropdown menu with "Implementing" selected.
 - Stem:** A text box containing the question text: "In your program you need to send a message to every person on the e-mail list. Which one of the following looping methods is most appropriate?"
 - Answer and Distractors:** A text box containing "do...while".
 - Difficulty:** Radio buttons for "very easy", "easy", "medium", and "hard".
 - Notes/Comments:** A text box for additional notes.
- Question Quality Checklist:** A table with columns for "Quality Item (click for explanation)", "Yes", "No", and "N/A".

Quality Item (click for explanation)	Yes	No	N/A
Clearly addresses the objective	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Has exactly one correct answer	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Uses novel content	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
Has only plausible choices	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>
- Question Preview:** A window showing the question text and answer choices:

Q: Applications of do...while loops
Q: In your program you need to send a message to every person on the e-mail list. Which one of the following looping methods is most appropriate?
A. do...while
B. for loop
C. while loop
- Template: Implementing:** A window providing detailed information about the template:

Q: Given task A, which of the following solutions is most suitable/likely?
A. [Solution based on applying procedure 1]
B. [Solution based on applying procedure 2]
C. [Solution based on applying procedure 3]
D. [Solution based on applying procedure 4]

Notes: Implementing occurs when a student selects and uses a procedure to perform an unfamiliar task. Because selection is required, students must possess an understanding of the type of problem encountered as well as the range of procedures that are available. Thus, implementing is used in conjunction with other cognitive process categories, such as Understand and Create.

Figure 5.4—Screen Design of the Question Entry Form

Figure 5.4 illustrates some more elements of the screen design that are focused on consciousness raising. Because there was so much information to include on this form, some means of allowing everything to be on screen at the same time was necessary. This is motivated by a usability goal of reducing the cognitive load on the teacher. (The

cognitive load refers to the amount of information that the teacher has to keep in their short term memory at any point in time. An example of this is when the teacher has to remember something on one screen while moving to another screen to get a related piece of information.) The screenshot shows the screen as it would look on a computer with a screen resolution of 1024x768 pixels. This is currently the most popular screen resolution in use. Statistics gathered from visitors to the NJIT IS Department website over the period between 10/2005 and 5/2006 indicate that 90% of visitors have a screen resolution at this level or above. Statistics from websites, such as www.w3schools.com indicate that in the general population the proportion of screens at this resolution may actually be between 70-80%. Designing QuesGen for screens at this resolution is somewhat risky, since the usability will be degraded for users with screens set at 800x600 or lower, but was a conscious tradeoff made in order to make the tool usable for what was considered to be the majority of users. The advantage is that the majority of users should now be able to keep the entire question writing on screen at one time.

A number of techniques were used to save screen space including tabbed windows, smaller fonts, and popup windows. The space for entering the answer choices allows the teacher to add up to nine distractors in addition to the correct answer choice. While it is unlikely that any question will actually have nine distractors, the capability is there. A problem with the tabbed text areas used for entering distractors is that you can't see all of the answer choices at once and it is harder to get a sense of how the question would look on a test. Clicking the "view preview" button displays a popup showing how the question would look as laid out for a test. The preview can be dragged around the screen so that it will not be in the way when accomplishing different tasks. For example,

while entering the answer choices the preview window could be moved to the right side of the screen over the checklist, but when it's time to use the checklist the preview could be dragged to the top or left while each checklist item is examined. Another technique for saving space was to use a smaller default font size. Figure 5.5 shows what the screen looks like when fonts have been enlarged. Although perhaps not as pleasing to the eye, the interface is still usable in this state. Some scrolling becomes necessary, but all the major information is still on the screen.

The screenshot shows the QuesGen web application interface. The main heading is "Add or Edit a Question". The form contains the following fields:

- Title:** Identify do...while scenario
- Objective:** Applications of do...while loops
- Template:** Implementing
- Stem:** In your program you need to send a message to every person on the e-mail list. Which one of the following looping methods is most appropriate?
- Answer and Distractors:** Answer: do...while; Distractors: D1, D2
- Difficulty:** very easy, easy, medium, hard, very hard
- Notes/Comments:** (empty)

On the right side, there is a "Question Quality Checklist" with the following items:

Quality Item (click for explanation)	Yes	No	N/A
Clearly addresses the objective	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has exactly one correct answer	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Uses novel content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has only plausible choices	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Distractors based on misconceptions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids trivial content	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is not culturally dependent	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Avoids opinion and questions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses humor cautiously	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Is grammatically correct	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses audience-appropriate vocabulary	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Uses parallel sentence structures	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Has clear instructions	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Writing is concise (not overly wordy)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

A "Question Preview" window is open, showing the question and answer choices:

Question Preview

Q: Applications of do...while loops

Q: In your program you need to send a message to every person on the e-mail list. Which one of the following looping methods is most appropriate?

A. do...while

B. for loop

C. while loop

Figure 5.5—The Question Entry Form with Enlarged Fonts

5.2.2 Counterconditioning

The second process of change that will be discussed, counterconditioning involves learning behaviors that can serve as a substitute for the unhealthy behaviors being

replaced. QuesGen identified the use of poorly-written MCQs and the use of MCQs in non-formative contexts as “unhealthy” behaviors that were targets for replacement.

Add or Edit a Question

Title: (REQUIRED) Best example of do...while loop

Objective: (OPTIONAL) please select an objective... ▼

Template: (OPTIONAL) Applications of do...while loops
Distinguish between control structures
Using switch and if statements

Stem: (REQUIRED)

Figure 5.6—Encouraging Use of Objectives

The suggested alternative behavior proposed by QuesGen is to use MCQs for frequent, informal assessments designed to help the students and teacher get a better sense for what concepts have been learned so far in a course. These are in line with the best practices for formative assessment described in Stiggins (2001), and Black and William (2003; 1998a; 1998b). The expression of counterconditioning is not explicitly found in any one bit of functionality, but rather pervades many of the features of the system. It also can be found in the lack of certain functionality found in other question writing software systems. For example, the use of multiple-multiple-choice items (i.e. students can select more than one of the answer choices) is not condoned by the research on effective MCQs. Because of this, it is not possible to create an item of this type with QuesGen, and such functionality will never be added. While not required, the inclusion of an explicit educational objective is strongly recommended by the system (see Figure 5.6), and the rationale for doing so is explained in the online tutorial on objectives. This suggestion is intended to replace the more common habit of teachers to leave the objectives of their questions unspecified.

5.2.3 Self-Reevaluation

The change process of self-reevaluation is described as journals or diaries, therapy, or other intervention strategies that involve some measure of introspection and examination

of one's current behavior. The goal is to take an honest look at current problem behaviors so that there is a clear understanding of where one should go next. Fishman et al. (2000) make it clear that opportunity to reflect upon one's teaching practice is a vital element of any successful professional development intervention. QuesGen has two specific features and one general feature that are designed to foster self-reevaluation. The form for entering educational objectives and the item analysis reports are the specific features, and the notes/comments spaces that appear in various places throughout the interface are another.

5.2.3.1 Educational Objectives

As a direct result of seeking to implement the tenets of formative assessment in his own teaching practice, the experimenter realized that the form for entering

Add or Edit an Objective

Title: (REQUIRED) Implement for loops in C++

Objective: (OPTIONAL) 3C. Apply Procedural Knowledge

Description: (REQUIRED) Students will demonstrate their understanding of loops and be able to implement them using correct C++ syntax.

Figure 5.7—Objectives Offer a Chance to Reflect

educational objectives is also an opportunity for self-reevaluation. Since beginning this research, the act of developing new questions and the requisite educational objectives to go with them has become an occasion to examine the deeper reasons why a given course is being taught. It forces the developer of the objective to examine not only the content of new questions, but also to go back to old questions and ask the sometimes difficult question of “what was this question really assessing anyway?” This can be a humbling experience and teaches the teacher to have an attitude of always seeking to understand exactly what the goals are of any educational setting.

5.2.3.2 Item Analysis Reports

Similar to the forms for filling out educational objectives, the item analysis reports are a feature which the instructors can use to examine their own teaching practice, particularly with respect to their assessment strategies. Reports such as this were incorporated into QuesGen with the intention that teachers' curiosity would lead them to look at them. The content of the report was designed to be presented in such a way that hopefully teachers will feel some measure of suspense waiting to see how the results of each new question turned out. These reports will hopefully lead teachers to either work to improve their question writing or to discard MCQs altogether from their assessment strategy. Either of these outcomes would be preferable from a pedagogical perspective to having them continue to use ineffective MCQs in a non-formative fashion.

5.2.4 Environmental Reevaluation

This involves an assessment of how one's behavior impacts the people and places one inhabits. It involves recognition that one's actions can serve as an influence or role model to others. It may involve rational or emotional processes. This process, perhaps more than any other, has great salience for teachers, whose profession it is to have an impact on the minds of their students. In some sense, the entire QuesGen tool is a method of environmental reevaluation. The chief means of accomplishing this process though is via the same functionality that was used for self-reevaluation: the item analysis reports.

5.2.4.1 Item Analysis Reports

The goal of the item analysis reports is to change teachers' knowledge and beliefs about how their quiz questions impact their students. For example, in the pilot study described earlier, some students expressed extreme frustration with some of the quiz questions where the answer choices were ambiguous, or where the answer choice in the course text conflicted with the answer given in the teacher's course notes. This frustration would have gone unnoticed by the instructor if this student hadn't been given a chance to provide feedback about the quiz questions. The impacts of these statements from the student on the teacher were significant. As recounted earlier, the teacher, who had decades of teaching experience, began to mistrust her own MCQs after that and became much more cautious in their application.

5.2.5 Self-Liberation

Self-Liberation is associated with what people refer to as "willpower," and actions toward self-liberation may take the form of promises or commitments, e.g. New Year's resolutions, to make certain changes in one's behavior. The chances of success increase when multiple options (three options seems optimal) for change are available. The perception of choice seems to make it easier for people to choose to keep their commitments. While it's difficult to refer to any of the features of QuesGen as self-liberation in the strict sense, there are two features—the ability to choose among templates, and the ability to create new templates—that provide the teacher with a means of choosing how he or she wanted to develop new questions with the tool.

5.2.5.1 Choice of Templates

The idea for question templates came from reading the revised manual on Bloom's Taxonomy of Educational objectives (Anderson and Krathwohl, 2001). In reading about each of the categories of objectives it became clear that there were certain formulas one could follow to develop a question that targeted a certain level of knowledge and skills in a reliable fashion. As shown in Figure 5.8, the ability for teachers to select among a set of pre-created templates allows them to be much more disciplined about the form of the questions that they author. To interpret this in transtheoretical model language, the teachers are given choices to make about implementing questions based on the needs of the students at the current time. In this way teachers are liberating themselves

from the need to rely completely on their own imaginations for new questions.

5.2.5.2 Custom Templates

The ability to create custom templates is already planned for phase two of QuesGen. While it will not be available in the first version, the idea of the template is powerful enough that soon teachers will want to begin developing their own, domain-specific

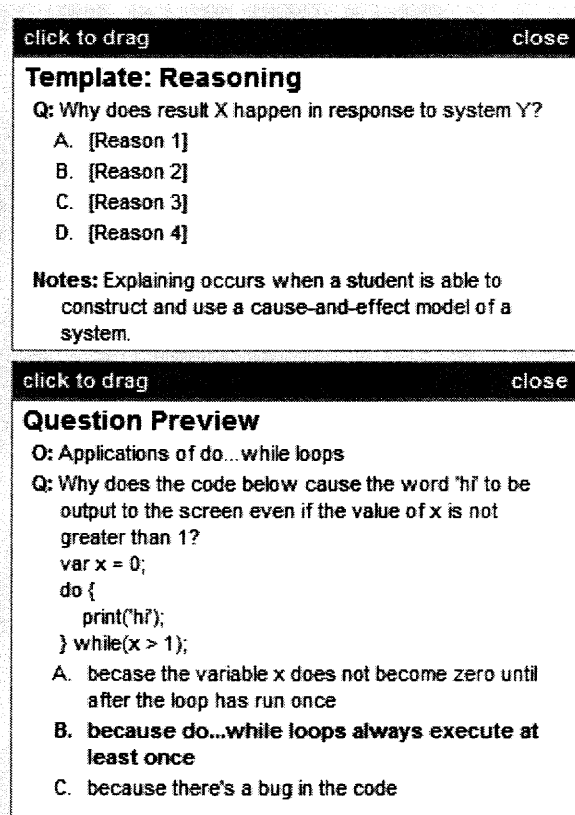


Figure 5.8—Using Templates for Reasoning Questions

templates—templates that are more specialized than the general ones that are delivered with the system. Energy invested in developing customized templates is a form of self-liberation because one is making a commitment to use the system for future question writing and also making a commitment to a higher standard of question generation for your students. These types of commitments are in line with what teachers are asked to do as part of high-quality professional development programs. They are asked to make a commitment to employ the new practices that they are learning.

5.2.6 Social Liberation

This process involves increased opportunities in social environments, such as smoke-free zones, salad bars, and other places where healthy behavior is socially promoted. The notion of social liberation has been stretched in this case to cover the interoperability features of QuesGen. The idea is that the healthy practices learned by using QuesGen can be taken out of the QuesGen environment and imported into other environments which did not previously support them, such as paper-based tests and other software packages like WebCT. As discussed in the more general software framework, the database schema used to represent the questions and objectives in QuesGen conform to IMS Global Consortium standards, and as such should be compatible with other learning management systems that also follow these standards.

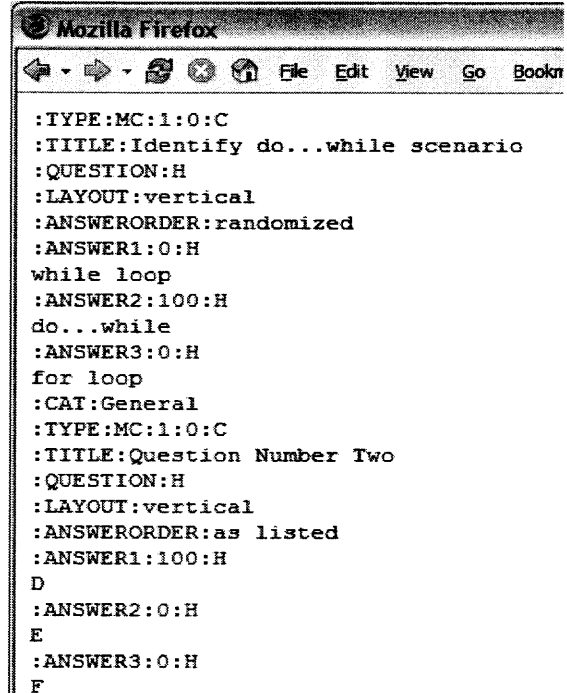
5.2.6.1 Printing Questions

This is a very straightforward function so no screenshot is generated to illustrate the functionality. When a teacher clicks “Print Questions” the questions in the database for that teacher are converted into a plain text format that can be saved ‘as is,’ or cut and

pasted into a word processor like MS Word. On pasting into Word, the formatting is preserved and the questions are ready to be printed and delivered to students.

5.2.6.2 Export for WebCT

If teachers want to use QuesGen within WebCT the tool can output the questions in the proprietary format necessary to be imported into the WebCT question database. At least one of the teachers in the second pilot study used this feature to import questions. The reverse functionality for importing WebCT questions into QuesGen has yet to be implemented, but is straightforward and on the list of things to implement for the next version. Interoperability is an important feature of all new software.



```

:TYPE:MC:1:0:C
:TITLE:Identify do...while scenario
:QUESTION:H
:LAYOUT:vertical
:ANSWERORDER:randomized
:ANSWER1:0:H
while loop
:ANSWER2:100:H
do...while
:ANSWER3:0:H
for loop
:CAT:General
:TYPE:MC:1:0:C
:TITLE:Question Number Two
:QUESTION:H
:LAYOUT:vertical
:ANSWERORDER:as listed
:ANSWER1:100:H
D
:ANSWER2:0:H
E
:ANSWER3:0:H
F

```

Figure 5.9—Export Questions for WebCT

5.3 Non-Implemented Processes of Change

For this version of the software it was not possible, nor considered wise to implement all of the processes of change. Since this is such a large framework, and since it is untested, it was felt that smaller elements should be developed and tested first before augmenting them further. However, this does not mean that other functionality was not considered

for inclusion. This section describes functionality that would be necessary to provide some electronic form of all of the remaining processes of change, but which was not implemented in this version of QuesGen. Each section is organized in the same way that earlier sections were discussed.

5.3.1 Helping Relationships

This involves enlisting the aid and support of close friends, family members, or advisors who can be supportive throughout the stages of change. In the context of teacher professional development, developing a community of practice is a well known feature of successful programs. Preferably the community that is learning and growing their teaching practice should have the opportunity to maintain those relationships over an extended period of time. There are at least two obvious sets of functionality that could be added to QuesGen to aid in this: online forums and a peer review system.

5.3.1.1 Online Forums

Online forums serve many communities as repositories of knowledge and wisdom about various topics. For communities based around a software tool, such as QuesGen, it is not uncommon to find that members of the community have provided extended examples, add-ons or plugins, or additional examples that clarify or augment the existing documentation of the system. If the QuesGen system turns out to be successful it may be worth the effort at some point in the future to make an attempt to seed a community around QuesGen use. Not only would this be a source of helping relationships, but also most likely a source of dramatic relief and social liberation as members of the community shared their stories and made commitments to practice formative assessment. Online

forums are rarely directly connected with the software in a meaningful way, but the next idea, the peer review system would tie the functionality for the forum directly into QuesGen in a meaningful way.

5.3.1.2 Peer Review

One of the recommendations of the research on question writing is that one should have a peer or colleague read and respond to questions before they are delivered to students. Doing a thoughtful review of a peer's questions is a task that takes considerable time and energy. Within QuesGen there might be a set of functions that allow teachers to submit their questions to the online community for review, and, in turn, to review questions that had been added to the community. Perhaps for every question one reviewed, one would earn the right to have one of his or her own questions reviewed. In this way one not only gets to have their questions reviewed, but also gets exposed to a great number of the questions developed by peers. Over time this is likely to deepen teachers' intuitive sense of which questions are high and low quality.

5.3.1.3 Question Exchange

A third form of helping relationship that is a natural extension of an online forum is a place to exchange questions. Similar to the question review, there might be some sort of point system so that in order to take questions from the bank, one must contribute questions, or pay money, which might be given directly to the question author. This addition to the QuesGen system would be an ambitious project on its own but would likely develop naturally as the online forum developed.

5.3.2 Stimulus Control

This process involves examining and removing temptations from the immediate surroundings. This is an interesting and unusual category to consider from the point of view of software design, but it is not difficult to imagine an environment which puts pressure on a teacher to write and use questions that are not of high quality and which are not employed in a formative fashion. Functions that fall into this category would have to bolster the teacher to stand against this outside pressure, or otherwise suggest to the teacher ways to get these pressures under control and maintain their commitment to writing only high-quality MCQs.

5.3.3 Contingency Management

When people engaged in behavioral change institute rewards and/or punishments in reaction to their efforts to change, they are using contingency management. At the level of an educational institution one could envision some sort of teacher recognition program for teachers who had made a commitment to improving their assessment ability and had followed through long enough to make a significant difference in their practice and in the learning of their students. It is not clear how such a program would translate into concrete functionality to be implemented within the QuesGen system, unless there was some sort of reporting system or submission system whereby teachers could submit their question statistics for comparison with other teachers. It is an interesting idea, perhaps, but will not likely be a priority in future versions of QuesGen.

5.4 Data Model and Scope of Functionality

This section lays out a more traditional view of the data model and scope of functionality. A diagram (Figure 5.10) of the elements of the QuesGen system will help clarify the scope of the functionality of the system. This diagram has been adapted from those found in two IMS published standards: the Question and Test Interoperability (QTI) specification, and the Reusable Definition of Competency or Educational Objective (RDCEO) specification.

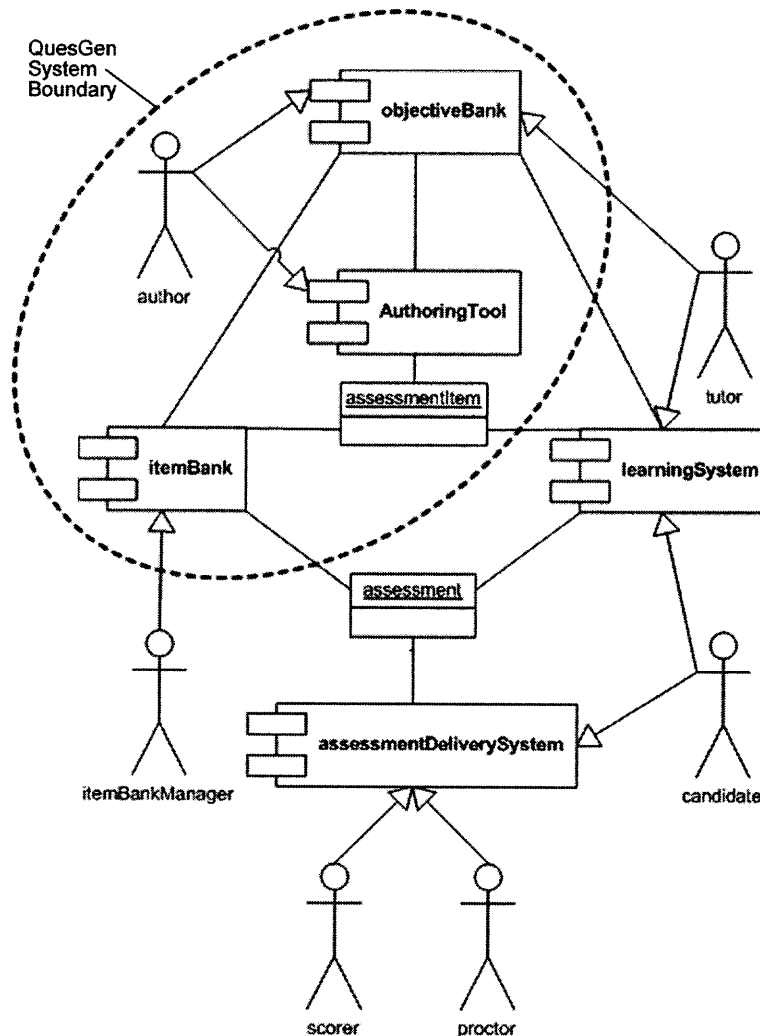


Figure 5.10—QuesGen System Boundary. Adapted from IMS QTI Overview, p5.

Figure 5.10 depicts a typical scenario in which the assessment of learning takes place. As this is a use case diagram, it should be understood that the actors shown represent different roles that a person might play in the assessment process. In most cases a single person may be responsible for more than one of these roles. Indeed it is possible to envision a self-study paradigm in which a single person would play all of the roles shown here. The dotted line encompasses the functionality proposed for the QuesGen system. The diagram helps emphasize the point that the focus of this research is on the “assessment item author” role that is frequently played by instructors. The diagram also indicates that the actor is an “author of educational objectives.” In practice, question authors are not always responsible for or have the prerogative to author, select, and/or change the educational objectives with which the questions they write must be aligned. As such the QuesGen system will be flexible enough to handle situations where the question author either does, or does not have responsibility for writing objectives. In both cases, it is necessary for the question writer to state explicitly which objective is being assessed with any given question.

It will be necessary that the questions developed with QuesGen be deliverable to students in some format in order to evaluate them thoroughly; however, the QuesGen system scope includes neither an LMS nor an assessment delivery system. Ideally, QuesGen could be plugged into one of the major LMSs on the market today, but it was decided not to pursue this option. One reason for this is that incorporation into a specific package would unnecessarily limit the pool of potential study participants to people who use that system. At NJIT, WebCT does not provide a programming API for the development of new plugins. While open source packages such as Sakai and Moodle do

offer such an API, they are not in broad enough use yet to merit the effort necessary to build plugins. Therefore a basic, web-based quiz delivery system was built to deliver quizzes. While this has the added benefit of allowing the post-quiz meta-questions (i.e. questions about the question asked to the students) to be incorporated into the actual quiz, the features of the question delivery system are not a part of the analysis for this study.

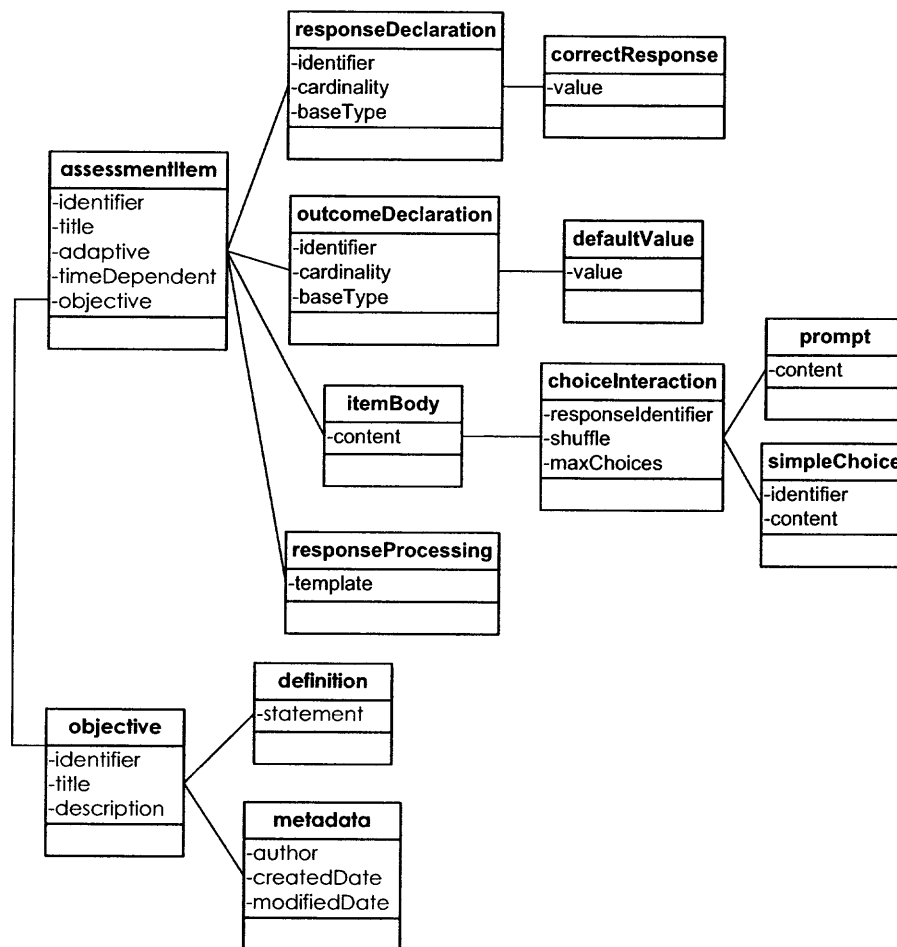


Figure 5.11—QuesGen IMS-compatible Data Model

The data model has been designed in compliance with IMS Global standards (www.imsglobal.org). Figure 5.11 shows a simple representation of the data model for the key components of the system. There are two primary objects: objectives and assessment items (i.e. questions). Every question must be associated with exactly one

objective, and in turn, each objective is expected to be associated with questions that assess it.

5.5 Conclusion

This chapter has presented the design of the QuesGen system, which was designed to partially implement a system model for the purpose of furthering teachers' professional development. QuesGen was designed to help teachers write better multiple-choice questions than they could without such a tool. The system was designed to inculcate an understanding of and a belief in the best practices for writing multiple-choice questions as set out by Haladyna. The processes of change in the transtheoretical model were used to guide the choices of which functions to create and how to prioritize them. The system design is also informed by relevant theories in information systems, namely captology, and adaptive structuration theory.

In the next chapter the concept of question quality is operationalized, and in the chapters that follow, an evaluation of this system is described that used university professors to determine whether or not the tool was successful in achieving its stated goal.

CHAPTER 6

OPERATIONALIZING QUESTION QUALITY

6.1 Introduction

This chapter discusses the operationalization of the concept of “question quality” and what it means for one MCQ to be “better” than another one. Since QuesGen’s stated goal is to help teachers write better MCQs, this issue is central to the evaluation of QuesGen’s success. This chapter begins in Section 6.3 with an overview and discussion of two quantitative approaches to determining question quality. Section 6.4 describes and discusses item-review panels which were used to score the questions on the quality of their writing and the depth of learning they measured. Section 6.5 discusses direct feedback from students which was gathered to balance and inform the ratings of the item-review panel. Section 6.6 describes how system usage logs were used to inform usage of QuesGen, and through it how other measures of question quality were interpreted. Section 6.7 describes the explanation building methodology and talks about how interviews with teachers were used to develop an explanation of why some questions were better than others. Section 6.8 briefly describes QUIS and UTAUT, two validated measures for learning about user satisfaction with an information system, and finally, 6.9 explores nuances in potential ways to define question quality in light of earlier discussion of the KBA Framework. The conclusion of this chapter will synthesize the discussion thus far of measures of question quality. The conclusion also introduces the following chapter, which will explain the methods by which QuesGen was actually evaluated.

6.2 What is a “good” question?

Put simply, a good question is one that helps an instructor learn the greatest amount about the current state of his or her students’ understanding both at the individual and aggregate levels. High quality questions inform the decision making processes of both teachers and students as they choose what and how to teach and learn. This definition is in keeping with the discussion of formative and summative assessment from Section 2.5. Summative assessment tends to serve needs or actors outside of the classroom such as credentialing, entrance examination, program auditing, or program performance evaluation. The results of summative assessments, by definition, are not used to identify, plan, or otherwise inform the instruction that goes on in a classroom following the assessment. Summative assessments may improve learning inasmuch as they are designed for overall program or curricular improvement; however they seldom if ever are directly beneficial to the students who take the actual tests. On the other hand, formative assessment, also known as assessment for learning, is primarily student focused. Formative assessments provide information to the instructor, the students, or both that can be used immediately to indicate whether a given topic has been mastered, or whether remediation is in order. At the class level, formative assessments can help an instructor make strategic choices about the formation of student teams, either pairing students with higher and lower levels of mastery, or grouping students according to mastery so that specific topics can be targeted directly to those students for whom they are most needed. The guiding question in the determination of measures of question quality, therefore, is “Does this measure provide insight into aspects of a question that have an immediate impact on the teacher-student relationship, and the student learning that results thereof?”

6.3 Quantitative Measures of Question Quality

The quantitative methods for determining question quality fall into two broad categories: classical test theory, and item response theory (IRT) (Hambleton, et al., 1991). Under classical test theory the key measure of question quality is known as the item discrimination index (DI). The one-, two-, and three-parameter logistic IRT models have been created by the professional test development community (e.g., the individuals who develop college entrance exams, etc.), and are currently the industry standard way of evaluating item quality. Both methods of analysis are described below. Following this description is a discussion of some of the underlying premises of these measures as well as commentary on their suitability in light of the previous focus on formative versus summative assessment. The DI was chosen to evaluate QuesGen. A discussion of the validity issues associated with selecting DI as the measure of question quality is included below.

6.3.1 The Item Discrimination Index (DI)

Item discrimination is described as the degree to which a question is able to discriminate between test takers of high and low ability. The formula for calculating discrimination is straightforward, and of the form:

$$D = (U_p - L_p) / U$$

where:

U = the number of students in the upper quartile

U_p = the number of students in the upper quartile who answered the item correctly

L_p = the number of students in the lower quartile who answered the item correctly

Discrimination values vary between -1 and 1 and are interpreted as follows:

- $D = 0$

The question does nothing to distinguish between test takers of high and low ability since all examinees in the upper and lower quartiles gave the same answer. This could be a result of a question that is either too easy or too difficult, but usually indicates an item that has problems and should be revised or discarded.

- $D = \pm 1$

The question perfectly distinguishes between high and low ability test takers. This occurs rarely and only when all high-ability examinees answer correctly and all low-ability examinees answer incorrectly (+1) or vice versa (-1).

- $D > 0$

High-ability test takers were more likely to answer correctly than low-ability test takers. This is a desirable outcome, and typically values of 0.30 or higher are considered acceptable, although the goal of question writers should be to maximize this value.

- $D < 0$

Low-ability test takers were more likely to answer correctly than those with high ability. Typically this means there is a problem with the question, such as that it is a trick question or worded in a way that is confusing.

The DI is used widely and comes built-in to the item analysis tools of some education software that support question development, such as WebCT. Despite this wide use, the item discrimination index (DI) has serious problems that bear discussion. The values returned for the DI on any given question are going to vary depending on the

group of students to whom the test is given, and also on what other questions are included on the test. To understand why this is the case, let's look again at the formula and its parameters:

$$D = (U_p - L_p) / U$$

where:

U = the number of students in the *upper quartile*

U_p = the number of students in the *upper quartile* who answered the item correctly

L_p = the number of students in the *lower quartile* who answered the item correctly.

Of importance here is the method by which the “upper” and “lower” quartiles are determined. In classical test analysis, the upper and lower quartiles are determined by ranking the test takers according to the total score on the test. This means that the discrimination score for any one question depends on how well students did on all the other questions. It also depends on how this particular set of students did on this particular set of questions. For example, if the ability level of the students was very homogeneous, it stands to reason that the DI for a given question would be lower since it is more difficult to distinguish between equally competent test takers. Indeed, if all examinees possess approximately equivalent ability, then on any given test, those who fell into the “upper” quartile and “lower” quartile may be a somewhat arbitrary categorization. The exact same question may have a higher DI with a group of students with heterogeneous levels of ability since it is easier to separate such students. Likewise, the DI will be sensitive to what other questions were asked on a given test to the extent

that these questions determine which students are classified into the upper and lower quartiles.

What this means, in practical terms, is that it is difficult to make reliable predictions about the performance of test items from one group of students to the next. Items might discriminate well with one group of students, and poorly with another. This state of affairs makes it problematic to use the DI as the sole measure of question quality. This idea is expressed by Hambleton et al. (1991) when they say:

Group-dependent items are of limited use when constructing tests for examinee populations that are dissimilar to the population of examinees with which the item indices were obtained. (p. 3)

The key here is that in order to make a prediction about how well an item will discriminate among examinees based on past performance, one must assume that the new set of examinees is similar to previous ones. This may not be an unreasonable assumption if the questions are delivered to students who have about the same amount of experience and instruction as past students who have taken the test, and who receive the test at about the same point in the sequence of instruction. Indeed, it is argued below that within QuesGen's intended context exactly these conditions will be present.

6.3.2 Logistic Models Derived from Item Response Theory (IRT)

The assumptions required to rely on the DI are not possible for modern psychometricians who are developing questions that will be seen by a wide variety of test takers, and may be used in a variety of contexts. Item response theory (IRT) grew out of the need to make reliable predictions about the performance of test items that are not group- or test-dependent. IRT has been successful in developing such models of item performance. One such model is the one-parameter logistic model, more commonly known as the

Rasch model after its developer. The Rasch model takes the form (Hambleton, et al., 1991):

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \quad i = 1, 2, \dots, n$$

where:

- $P_i(\theta)$ = the probability that a randomly chosen examinee with ability θ answers item i correctly
- b_i = the item i difficulty parameter
- n = the number of items in the test
- e = a transcendental number (like π) whose value is 2.718 (correct to three decimal places)
- $P_i(\theta)$ = an S-shaped curve with values between 0 and 1 over the ability scale

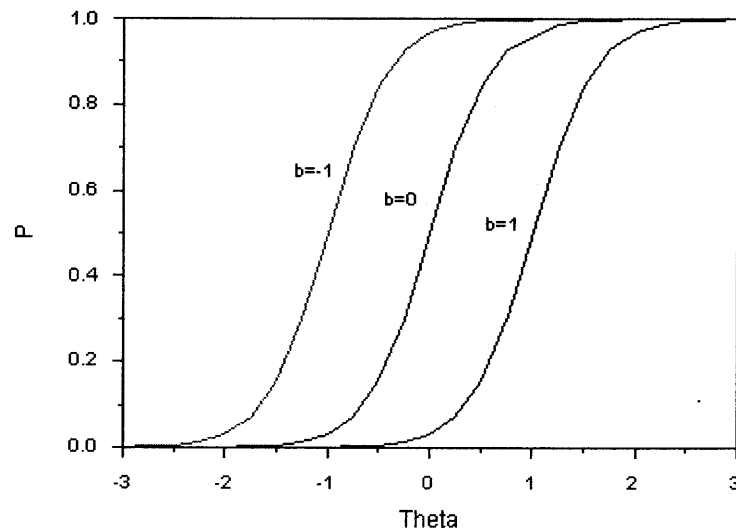


Figure 6.1—Sample Item Characteristic Curves (ICC)
(image from <http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm>)

Figure 6.1 shows a set of sample item characteristic curves that might be obtained using the Rasch model. One curve represents one question. The horizontal axis is the ability scale, and the vertical axis is the probability of a correct answer. Hence, as one

moves from left to right, i.e. from lower ability to higher ability, the probability of a correct answer increases. The slope of the curve represents the degree to which the item discriminates between high and low ability examinees. A vertical line would perfectly discriminate between examinees with θ values to the left and the right of the line. A horizontal line would indicate zero ability to discriminate since all examinees would have an equal chance to answer correctly. A negative slope would indicate a problem question since lower ability examinees would have a higher chance of answering correctly. The ability scale, θ , is an absolute scale, and therefore the ability of a question to discriminate between ability levels will be the same regardless of the characteristics of the examinee, or the test context in which the question is delivered. The difficulty of a question is measured by how far to the left or right the curve is shifted. In the figure, the middle curve ($b=0$) denotes a question where an average examinee ($\theta = 0$) has a 50% chance of answering correctly. The curve to the right is more difficult since students with ability one standard deviation above normal ($\theta = 1$) have a 50% chance of answering correctly. Likewise the curve to the left represents an easier question.

The Rasch model is a 1-parameter logistic model, but the 3-parameter item response function is more popular among professional testing organizations due to the increased amount of information that can be learned from it. The 3-parameter model takes the form:

$$P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad i = 1, 2, \dots, n$$

where:

$P_i(\theta)$ = the probability that a randomly chosen examinee with ability θ answers item i correctly

- D = a scaling factor introduced to make the logistic function as close as possible to the normal ogive
 a_i = the item i discrimination parameter
 b_i = the item i difficulty parameter
 c_i = the item i pseudo-chance level parameter
 n = the number of items in the test
 e = a transcendental number (like π) whose value is 2.718 (correct to three decimal places)
 $P_i(\theta)$ = an S-shaped curve with values between 0 and 1 over the ability scale

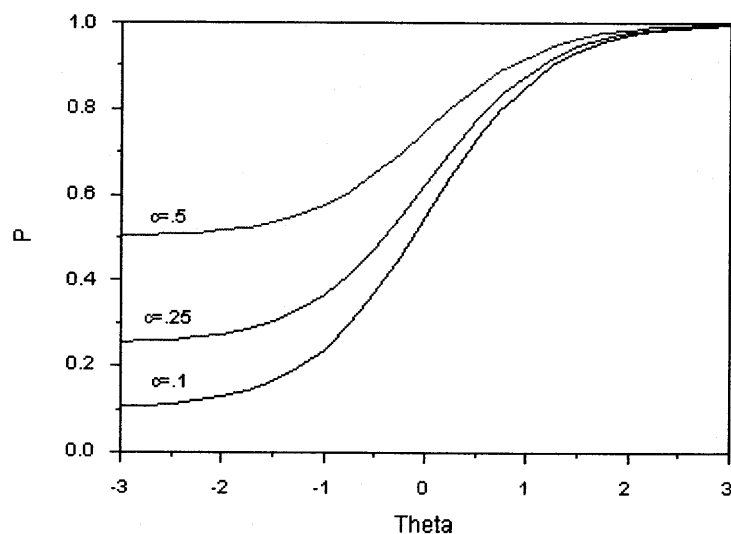


Figure 6.2—Sample ICC Curves for the 3-parameter Logistic Model
 (image from <http://luna.cas.usf.edu/~mbrannic/files/pmet/irt.htm>)

Figure 6.2 shows three sample ICC curves for the 3-parameter logistic model. Again the horizontal axis is the ability scale, and the vertical axis represents the probability between 0 and 1 of answering correctly. This model is a generalization of the Rasch model and includes two parameters not in that model. The parameter a is the discrimination parameter and is proportional to the slope of the curve for any given value of ability, θ . For well-specified questions, the a parameter takes values in the range $(0, 2)$.

Higher values of a indicate a steeper ICC and hence a greater ability to discriminate between test takers at a given ability level. The c parameter, known as the *pseudo-chance level* parameter, takes into account the notion that even an examinee with zero ability still has a chance to answer a multiple-choice item correctly by guessing. The intercept with the vertical axis represents the c parameter. In the picture the topmost curve shows a 50% chance of guessing correctly even if the examinee has zero ability. True/false questions might display behavior such as this.

While using a model such as the Rasch model to measure question quality would clearly be preferable to DI, such models do not come without their own complications. The chief problem in the context of the QuesGen system is that neither the ability parameter, θ , nor the difficulty parameter, b , can be calculated or known *a priori* for any given question. These parameters must be estimated using a known method of statistical parameter estimation. Maximum likelihood estimation (MLE) is frequently chosen for this task, but unfortunately MLE requires having at least 200 examinee responses before parameters can be calculated with any accuracy for Rasch models; roughly 1000 responses are necessary to accurately estimate all of the parameters in a 3-parameter logistic model (Parshall, et al., 1997). This is feasible for professional test development companies which usually have an “experimental” section on most tests where new items are debuted in order to get the necessary data for parameter estimation. For the typical classroom instructor, however, it’s not likely that access to such great numbers of examinees will be available. A further discussion of DI and IRT’s appropriateness as measures of question quality for evaluating QuesGen follows.

6.3.3 Discussion of DI, IRT, and QuesGen

Are DI and IRT appropriate measures of question quality to use in the evaluation of QuesGen? QuesGen seeks to improve the quality of MCQs that teachers write. Both DI and IRT measure a question's ability to categorize students into low and high ability groups based upon mastery of the concepts assessed by a given question. A crucial question therefore is whether or not "discriminatory power" is what we mean when we refer to "quality." This broader, more philosophical question will be revisited at the end of this chapter. In the meantime, given the context in which QuesGen is designed to be used, it is argued that DI is a more practical and useful measure despite its shortcomings of test and group dependency.

The choice to reject IRT logistic models as the empirical measures of question quality for use in QuesGen was made primarily based on the number of examinees necessary to make such models function. Some large university lecture courses may have upwards of 200 students, but this is not the norm and should not be a requirement for using QuesGen effectively. Future versions of the tool may make IRT-based analysis available for teachers who meet its requirements, but for now, the choice was made to use the discrimination index (DI) as the empirical measure of question quality to be used in QuesGen. In order to do so, DI's shortcomings must be addressed.

First we will address DI's external validity. It is important to draw an important contextual distinction between the context of IRT/DI literature and QuesGen's target context. The majority of the assessment literature is targeted at professional test development situations where items are used for sensitive and high-stakes achievement and credentialing examinations. In those cases, when a question fails it can have serious

ramifications for the examinees, such as failure to gain entry to college, or failure to get a needed professional certification. “Failure” of a question means that a person who is, in truth, qualified is rejected because of a mistake in the form or scoring of that question. Such failures increasingly have legal ramifications for the organization responsible for developing the test. QuesGen, on the other hand is designed specifically for low-stakes assessment in classroom settings. The goal of QuesGen is not to rank students more accurately, nor to make decisions about who should or should not be admitted to competitive programs. QuesGen is designed to help instructors gain a clearer sense of what their students have and have not mastered from the content of a course. Since the DI is designed specifically to allow instructors to discriminate between students of high and low ability, it is externally valid.

Given that the DI can be argued to be a representative measure of question quality, its internal validity must next be addressed. A major weakness of classical methods of item analysis is that the indices produced with these methods tend to be very group-dependent. In other words, if the same indices were generated with the same questions on the same test, but on a different set of examinees, dissimilar from the first, the indices are likely as not to be significantly different. For example, it would not be appropriate to use a math question developed for third-graders on an achievement test for sixth graders. One must exhibit caution in taking a question and making a general claim that it will have equal discriminatory power regardless of the context in which it is delivered. Fortunately for QuesGen, however, it is not likely that questions will be used in a variety of different contexts due to the nature of classroom teaching.

It is argued that characteristics of classroom teaching lessen group-dependent shortcomings of classical test theory. Classroom teachers typically teach the same subjects to students at the same level from year to year. Students coming into these classrooms typically have similar characteristics such as all having earned admission to the same university, coming from roughly the same age group, having taken the same pre-requisite courses, being in a similar major, and having similar career goals. As such, a reasonable assumption is that for the purposes of classroom assessment it is relatively safe to ignore the group-dependent weaknesses of DI. The primary reason for this is that when teachers develop questions, they develop them with a specific set of students in mind, for use at a specific point in the course of instruction. Different from professional test developers who do not know, and will likely never meet, the examinees who answer their questions, classroom teachers are more likely to have a personal relationship with examinees. Research indicates that despite this personal contact, teachers are not good at monitoring what their students actually understand (Chi, et al., 2004). QuesGen is designed to make them more proficient at this skill by making it easier for them to develop frequent formative assessments that will gauge student mastery.

Apart from group-dependence, the other major shortcoming of DI is test-dependence. Indices generated under a classical test model will change depending upon the specific set of questions that are included in a given test. To the extent that the other questions on a test change the determination of which students are in the “upper” and “lower” quartiles, this will also change the ability of any individual question to discriminate between students of “high” and “low” ability. Test-dependence will not be a factor in the evaluation of QuesGen because no attempt will be made to use or compare

questions when they are used on different forms of a test, in different combinations; question quality evaluations will all be made within the context of a single test.

Can the DI be used to compare questions written by different instructors, used on different tests, for different sets of students? Given the above discussion it seems reasonable for instructors to use the DI to evaluate their own questions in specific instances, but for the purposes of evaluating QuesGen's effectiveness it is important to address the issue of the DI's suitability for cross-classroom comparison. Phrased another way, is it meaningful to say that a calculus question with a DI of 0.532 is "better" than a history question with a DI of 0.377? What assumptions must be made to make this a valid comparison? The next paragraphs will argue that the assumptions necessary to make cross-classroom comparison of questions possible are that one, the courses should be the same difficulty level, and two, the students in the classrooms be randomly sampled from the same underlying population of students.

What limitations are imposed by the shortcomings of the DI discussed above? The DI is sensitive to two things: the group of students that take an exam, and the set of questions that accompany a particular question on the exam. The chief limitation imposed by this sensitivity is that one cannot take a question out of the context of a given set of questions delivered to a given set of students and *predict* what the DI of the question will be if delivered to a different set of students accompanied by a different set of questions on a different exam. In other words, DI does not provide the confidence necessary to *reuse* questions in dissimilar contexts. Reuse is a primary goal of the professional testing industry, but arguably not the primary goal of classroom instructors. In the context of evaluating QuesGen, the goal of calculating DI is to make a spot

judgment of the quality of a question in a specific instance. Since this is a different context than is discussed in the professional testing literature, it is necessary to ask if there is anything about the group- and test-dependent sensitivities of DI that would make it unsuitable as a basis for comparison.

A first question is, does course content bias the value of DI—i.e. is there something about calculus, for example, that would make it consistently more likely to produce DI scores that are higher or lower than DI scores calculated for history content? Likewise, is there something about upper-level calculus that makes its questions consistently more or less likely to have a higher DI than lower-level calculus? It is argued that while course content, per se, does not bias DI, variance in the *difficulty* of course content may have an impact. For history questions and for calculus questions alike, since a classroom instructor writes a particular set of questions to assess mastery of the content that he or she has taught in the classroom, it is reasonable to assume that all other things being equal, there should be no difference in the ability of the questions to discriminate between students who have or have not mastered the content in question. This reasoning would support the use of DI to do cross-classroom comparison. However, if the content of the course is particularly easy or particularly difficult, this will tend to reduce the variance in student mastery—i.e. it will be more difficult to distinguish differences in student mastery in a course where the material is especially easy (since all students will master it), or especially difficult (since all students will fail to master it). Hence, one would expect lower DI scores for courses that were particularly difficult or easy.

Therefore, a necessary assumption required to compare DI across different courses is that the difficulty level is approximately the same. In practice it may be difficult to make a meaningful comparison between the difficulty of courses in two different domains, e.g. calculus and history. On the other hand, such comparisons are made routinely when instructors in colleges refer to courses as “freshman level” or as “300-level” as opposed to “100-level” or “400-level.” Indeed, for the evaluation of QuesGen, such a heuristic will be used to argue for the meaningfulness of cross-domain comparisons of DI scores. At the same time, it is recognized that the lack of direct comparability between the difficulty of courses weakens the DI as an absolute measure of question quality.

A second factor that may bias DI scores and threaten the ability to use them for cross-course comparison of question quality is variance in the ability levels of the student population in the courses to be compared. For example, the students in an “honors” course are typically pre-screened based on past high achievement. It is reasonable to assume that the variance in ability level in the honors course is less than the variance in ability levels of students in a non-honors section. Therefore it would be more difficult to write a question that discriminates between students of “high” and “low” ability in the honors course, than it would be in a regular course, resulting in lower DI scores in the honors course. These lower DI scores may or may not reflect questions of higher or lower quality.

As such, a second assumption required to compare DI across different courses is that the students enrolled in those courses all be randomly sampled from the same underlying population of students. Differences may arise from students being in different

majors, different years (upper versus lower classmen), or from different universities. For the purposes of evaluating QuesGen, it will be necessary to ensure that such differences are controlled to allow for cross-course comparison of DI scores.

In summary, it is argued that DI is an appropriate measure of question quality for use in the evaluation of QuesGen for three reasons. First, it is readily available, i.e. easily calculated from student quiz scores. Second, since the goal of questions in a classroom setting is to discriminate between which students have and have not mastered course content, a “discrimination index” is an externally valid measure of the quality of these questions. Third, the DI can represent an internally valid measure of question quality *if* certain variables surrounding the use of the questions are controlled. If one can control for the relative difficulty of courses and for the homogeneity of variance of ability levels of the students enrolled in the courses, it is reasonable to use DI to compare questions across different courses. Meeting these requirements is challenging but not impossible. These requirements will be addressed in the study design to be described in the next chapter. The DI is not the only measure by which question quality was assessed—in the next section item-review panels are discussed.

6.4 Item-Review Panels

In addition to statistical measures such as DI and IRT, the professional test industry also makes heavy use of expert review (Engelhard et al. 1999, Haladyna 2001). Typically a panel of two to three subject-matter experts (SME) is convened, and each panel member is given a set of questions to review along with a standard set of review criteria. Items are reviewed to determine whether or not they contain any significant technical, factual,

and/or cultural flaws. Where flaws are found, questions are typically sent back for revisions or discarded altogether. For the purposes of evaluating QuesGen's effectiveness, an item-review panel was convened to review the quality of the questions generated with the system. This section discusses the item-review instrument that was used by the panel to determine question quality.

The criteria for evaluating question quality were captured on an item-review instrument (Appendix C) which was designed using the empirically validated features of high-quality questions reviewed in section 2.6 and included in the question-quality checklist feature of the QuesGen system described in section 5.2.1.3. One concern was that since the criteria on the item-review instrument and on the question quality checklist were the same, an evaluation that rated the quality of the questions based on these criteria would be perceived as tautological. There is no simple way around this issue. It would be counterintuitive to change the list of criteria on either the checklist or the review instrument for the sake of avoiding a tautology when the criteria are so clearly indicated by the available empirical evidence on what makes a question a good question. One observation about information systems in general, and one change to the instrument were used to justify the validity of the instrument for use in evaluating QuesGen.

First, it was observed that the mere inclusion of a feature in a system has very little to do with whether or not users of the system will actually use that feature. It is very well known that people who use a software system tend to ignore system features that are unfamiliar, particularly when their utility is not apparent. Indeed, it is arguable that one of the major foci of all of IS research is figuring out how to get people to use the features of a system that have been carefully designed to increase their efficiency and/or

effectiveness at completing a task. In that vein, it was predictable that even though the checklist was available to all of the people who used the QuesGen system, many would not spend time looking at it or incorporating the recommendations of the checklist into the writing of their questions. Not only that, but system users were not expected to be familiar with all items on the checklist. As such, it would require extra time and effort for QuesGen users to become familiar with the meaning of all of the checklist items. Since it could not be expected that all users would put forth the effort required to learn about the checklist items, this seemed to lessen the possibility of a tautological relationship between QuesGen users' adherence to the best practices captured on the checklist and the item-review panel's likelihood of perceiving this adherence.

Second, a change was made to the item-review instrument to add some depth and nuance to its application and interpretation. Whereas on the question quality checklist each of the checklist items represented a dichotomous choice (i.e. the checklist item was observed—yes/no), the selection on the item-review instrument was expanded to represent a 5-choice Likert-type item (strongly disagree to strongly agree). These extra choices were added to allow the item reviewers to bring their finer sensitivities to bear when making an evaluation of the question. It was also observed that the degree to which the judges and QuesGen users agreed about the extent to which the checklist items had been followed could be measured.

In summary, item-review panels are a fixture of item-analysis in professional testing situations, and as such represent an externally valid measure of question quality. A trained human eye is able to examine each question for flaws. The item-review instrument employed is made up of items that come from empirically validated research

on MCQ quality, and as such represents an internally valid measure of question quality. It is arguable that using the DI and item-review panels is sufficient to gauge question quality, however, several other measures are discussed below that were included to bring greater depth and nuance to the analysis of QuesGen including direct feedback from students, interviews with teachers, and satisfaction measures with the system.

6.5 Direct Feedback from Students

In the pilot study (Benton, et al., 2004), it was found that the qualitative comments made by students yielded useful information as to why students missed questions. Frequently the reason students missed questions was through no fault of their own, but because of problems with the questions. Student feedback was once again solicited from students on the questions written with QuesGen. As in the pilot, the survey gathered students' opinions of each question's difficulty, clarity, fairness, vocabulary, and, if they missed the question, the reason why. Correlations of student feedback with the responses of the item-review panel and with the DI scores were expected to increase understanding of why questions performed in the way that they did.

As discussed earlier, because of experience gained in the pilot, cheating is a concern. There exists no simple way to make online quizzes secure other than restricting the access to the quiz to a short time frame, say one hour, and requiring the students to take the quiz in a location where they can be monitored, such as a classroom. In order to discourage cheating, instructors were encouraged to remove the incentive to cheat by making the quiz an optional or possibly extra credit exercise. To provide additional incentive for students to complete the quiz, given that they would receive little or no

course credit, students who completed the quiz and survey were entered into a raffle for a \$25 gift certificate.

6.6 System Usage Logs

An important piece of the information gathered to assess question quality was teachers' actual use of the system. To that end, each time a page was requested this was recorded in the QuesGen system usage logs. The log also recorded the user that made the request. This allowed for an analysis of what features were used, how often they were used, for how long, and by whom. It was possible to monitor how many times users logged in, whether or not they watched the videos and tutorials, and whether or not they used QuesGen features such as the question quality checklist, question templates, and question objectives. Analysis of the usage logs allowed correlation analysis to be completed between actual system usage and the various measures of question quality. It also allowed some usability issues to be uncovered. Finally, system logs were used to tailor the questions used in the follow-up interviews conducted with teachers, which are discussed next.

6.7 Interviewing the Teachers—Item Review Sessions

The item review sessions with teachers were designed to accomplish two things. First, there was a desire to thank instructors for their participation and provide them with tangible feedback that they could use to improve their own assessment practice in the future. Second, there was the desire to have an opportunity to ask any item-specific follow-up questions that existed about their quizzes and also to get semi-structured

feedback about their experience with the system, their desire to use it in the future, and any suggestions they had for its improvement. The procedure followed for these interviews was as follows.

After students had taken their quizzes and provided feedback, and after the expert judges had reviewed all of the questions, a question analysis report (QAR—see sample in Appendix D) was generated for each instructor. The QAR was about twenty-one pages long for each instructor and had about two pages of analysis for each of the questions that the instructor had written. The QAR displayed the following:

- The title of each question
- The objective that the instructor had selected to be associated with that question if it existed (QuesGen interface only)
- The question template upon which the instructor had based the question (if selected—QuesGen interface only)
- The actual question
- The number and percentage of students who selected each of the answer choices including the correct choice
- The student responses to the post-quiz survey—i.e. students' indications of the question's difficulty, fairness, and clarity, words they didn't understand, their reasons for missing the question, and other comments they had about the question
- The objectives from the lecture along with the expert judges' estimations of the extent to which the question addressed each of the objectives
- A list of the best practices which the judges agreed had been followed for this question

- A list of the best practices that the judges agreed this question had violated
- The judges' assessment of where the question fell in Bloom's taxonomy
- The discrimination index of the question.

After the QARs had been created for all of the instructors, the instructors were contacted and an individual interview was scheduled with each person. The interviews were designed to take between 45 minutes and one hour (see interview guide in Appendix E). After obtaining consent to video tape the interviews, the interviews were conducted. The interview was divided into two parts. In the first part, the instructors were asked about their teaching backgrounds, their attitudes toward students, toward the course they were teaching, towards assessment in general and MCQs in particular, their experience with and use of educational objectives, and then each instructor was asked about their impression of the video lecture that was prepared for use in the QuesGen study. In part two of the interview, the interviewer explained the QAR in detail and reviewed several of the questions that the instructor had written with the instructor. The instructor's feedback was solicited about the question with respect to the indications of quality that were reported on the QAR.

The interviews were used as part of a research strategy called "explanation building."

Under explanation-building, the researcher does not start out with a theory to be investigated. Rather, the researcher attempts to induce theory from case examples chosen to represent diversity on some dependent variable (ex., cities with different outcomes on reducing welfare rolls). A list of possible causes of the dependent variable is constructed through literature review and brainstorming, and information is gathered on each cause for each selected case. The researcher then inventories causal attributes which are common to all cases, common only to cases high on the dependent variable, and common only to cases low on the dependent variable. The researcher comes to a provisional conclusion that

the differentiating attributes are the significant causes, while those common to all cases are not. Explanation-building is particularly compelling when there are plausible rival explanations which can be rebutted by this method. (Garson, online, accessed 4/11/07)

Explanation building is a methodology that builds its credibility and external validity because the goal of the interviewer is to falsify the working hypothesis. As successive interviews fail to provide the evidence needed to falsify the hypothesis (*if that is what actually happens*), a case is built supporting the hypothesis. In this approach, it is common that the original hypotheses are somewhat provisional and designed to be refined as the process continues. The interviews were designed to provide evidence to answer the research questions focused on teachers' attitudes towards students, assessment and MCQs.

6.8 User Satisfaction

To gauge teacher satisfaction with the tool, two pre-validated instruments were used: the Questionnaire for User Interaction Satisfaction (QUIS), published by the University of Maryland's Human-Computer Interaction Lab (HCIL) (Chin, et al., 1988), and the Unified Theory of Acceptance and Use of Technology (UTAUT) published by Venkatesh, et al. (2003). The surveys were built into the QuesGen interface, and a link to them was made active after the teachers had completed writing their questions and delivering their quizzes to their students. Significant flaws in the interface design could have a serious impact on teachers' ability to complete the tasks for which QuesGen was designed and therefore a significant impact on how the other measures described above should be interpreted. The results of the surveys were used to help interpret the other results more

directly related to question quality and in the improvement of the user interface. The surveys and their usage are described below.

QUIS gathers information on users' impressions in nine categories: system experience, overall reactions, screen design, system terminology, learning, system capabilities, online help, online tutorials, and multimedia. Left out of the questionnaire were the sections on teleconferencing and software installation, neither of which was applicable to the experience of QuesGen. QUIS was expected to yield information about aspects of the user experience that are not directly related to the task of writing MCQs. In addition to specific recommendations of user interface improvements, cases of extreme difficulty with the interface were identified and in one case indicated that a user's quiz data should be removed from further analysis of the tool.

UTAUT does not focus on the specific features of a system, but instead aims to shed light on general factors that affect whether or not people are likely to use a system in the future. The items of the UTAUT instrument were adapted to apply to QuesGen (Appendix I). UTAUT measures user attitudes on six factors: performance expectancy, effort expectancy, facilitating conditions, social influence, behavioral intention to use the system, and voluntariness. The goal of collecting this information was to provide added depth and insight into the interpretation of the other measures of question quality discussed above. For example, would there be a correlation between question quality and instructors behavioral intention to use the system in the future.

This completes the description of the measures of question quality that were gathered towards the goal of evaluating QuesGen. Before concluding this chapter, other interpretations of question "quality" that came out of this analysis are discussed.

6.9 Other Interpretations of “Quality”

QuesGen’s stated goal is to help teachers write better multiple-choice questions. Chapter 4 laid out a model for the building of a class of information systems deemed likely to be able to encourage teachers to adopt best practices that would change their underlying attitudes toward assessment and thereby increase their question-writing ability. While in large part the design of QuesGen is based upon the model laid out in Chapter 4, QuesGen is not robust enough at this stage to allow a validation of the soundness of the model. While the validation of the model is one of the objectives planned for QuesGen in the future, such will not be attempted in this thesis. However, consideration of the model gives rise to interesting alternative interpretations of the meaning of a “high quality” MCQ.

Table 6.1 categorizes the measures that were considered for inclusion in this thesis. It should be clear that the measures in the TTM category were not chosen. The rest of this section will discuss the measures that were *not* chosen, and begins by exposing a subtle, but significant nuance in the goals of QuesGen.

Question Quality	KBA → Behavior
<u>Empirical Measures</u> <ul style="list-style-type: none"> • Item Discrimination Index (DI) • IRT-based measures <u>Qualitative Measures</u> <ul style="list-style-type: none"> • Question Quality Checklist • Teacher interviews • Student feedback 	<u>Transtheoretical Model Based Measures</u> <ul style="list-style-type: none"> • Decisional balance • Self-efficacy • Temptation <u>Behavioral Measures</u> <ul style="list-style-type: none"> • Type/frequency of tool usage

Table 6.1—Potential Measures of QuesGen's Effectiveness

The nuance is this: when QuesGen purports to “help teachers write better multiple-choice questions,” it could mean two things:

1. “Better” could represent an objective quality of the artifacts (i.e. MCQs) developed by the teachers at any give point in time, or
2. “Better” could represent an underlying KBA which signifies a teacher’s ongoing commitment to producing higher-quality MCQs, and which may translate over time into the actual production of objectively better MCQs.

The rest of this section explores what it would mean if the second meaning of quality were chosen.

The literature review, and the subsequent systems model, defended a position that the way to change teachers’ behavior was to change their underlying Knowledge, Beliefs, and Attitudes (KBA). The assumption is that if teachers possess KBA which is aligned with the best practices (as defined by Haladyna) for writing and using MCQs, then they will, in fact, write high-quality MCQs. Therefore, an indirect way to measure whether or not teachers are able to write better MCQs is to assess their KBA. Pursuing changes in KBA is desirable from a professional development standpoint because, although they are harder to change, once changed, KBA tend to be relatively stable properties of an individual. Therefore, if a strong KBA is developed in a teacher, that teacher will help foster a significantly higher amount of learning in students over the long term.

The software design model provides a set of metrics, based on TTM, that efficiently capture information about KBA: decisional balance, self-efficacy, and situational temptation. Although the diagnostic information that these measures provide is standard, and its interpretation uniform within TTM, the actual content of the measures is completely different depending on the behavior to which TTM principles are being applied. To give a simple example using decisional balance (again, decisional balance

for a target behavior refers to the difference obtained from subtracting the number of cons from the number of pros that a person is aware of at a given point in time), the decisional balance for smoking cessation (i.e. awareness of pros and cons for quitting smoking), is completely different than the decisional balance for exercising (i.e. the pros and cons of adopting an exercise regimen). This means that for a new behavior, such as the formative use of high-quality MCQs, these measures must be developed from scratch and validated.

TTM-based metrics of KBA are not the only ways to measure behavior, since behavior can be observed either directly or indirectly. It is possible through observables such as system logs, the actual artifacts of use (i.e. questions), diaries, interviews, or other means, to measure a teacher's actual behavior with respect to the development and use of MCQs. Indeed, a good deal of this information will be collected in the process of evaluating QuesGen. However, using the second meaning of "better" in the nuanced definition above, it is not likely that any type or amount of use of QuesGen would allow the conclusion to be made that a teacher had adopted the KBA aligned with effective use of MCQs. In a normal school or university setting, if a teacher used QuesGen to develop high-quality MCQs, this may be a form of evidence that the teacher had acquired the target KBA. But the crucial difference between that scenario and this research is *choice*. In a normal setting, a teacher would have a choice of whether or not to use QuesGen, and so use *may* indicate adoption of target KBA. However, in this study, teachers were asked to use QuesGen and so their behaviors were not strictly voluntary and didn't represent the actions of people who have acquired a new behavioral pattern.

For this thesis the decision was made to restrict the meaning of "better" to the objective observation of the quality of the actual MCQs developed by the teachers. It is

expected that studies in the near future will be able to capture *optional* use of QuesGen, and therefore have some potential to make conclusions about teachers' underlying KBA, but this is a step that must come after the first step, which is measuring outcomes when teachers are explicitly *requested* to use QuesGen, as will be done in this study.

6.10 Summary

If the goal of QuesGen is to help teachers write better MCQs then the definition of what “better” means is crucial to a meaningful evaluation of the system. In this chapter, a number of ways of operationalizing the concept of “question quality” have been discussed. To synthesize them all here, a “better” question will exhibit the following:

- Relatively higher ability to discriminate between test-takers of higher and lower ability as measured by the discrimination index (DI)
- A greater tendency to be aligned with explicitly stated educational objectives
- Less likelihood of violating the best practices of MCQ construction
- A higher chance of being classified as “higher order” in Bloom’s taxonomy

And in the context of the QuesGen system, these factors may be influenced or explained in part by paying attention to a number of other factors including:

- How, and how much instructors used the system for writing the questions
- Teachers’ description of the context surrounding the time when they wrote the questions

In the next chapter, a detailed description of the study designed to evaluate QuesGen and implementing the measures described here follows.

CHAPTER 7

QUESGEN EVALUATION

7.1 Introduction

In the first chapter of this thesis, the motivation for building QuesGen was introduced. In Chapter 2, the literature and relevant research provided insight into how QuesGen might be built. This was further developed in Chapter 4 which described an entire class of systems designed to bring about teacher change. Chapter 3 documented some first steps in the research process. Chapter 5 described the actual tool that was built, and Chapter 6 explored methods for evaluating the tool. This chapter describes how QuesGen was evaluated and builds a theoretically based case for the expected outcomes of QuesGen's use. It starts out by describing and explaining the hypotheses. After that follows a detailed narrative of how data were collected. The narrative highlights key choices and experiment design decisions that were made to control variance. This chapter lays the groundwork for Chapter 8, which will present the data collected and the data analysis.

7.2 Hypotheses

In this section the research questions that were asked in Chapter 1 will be revisited, and hypotheses related to QuesGen's impact developed.

7.2.1 New Functionality

The first set of research questions in Chapter 1 asked whether or not the inclusion of the new types of functionality proposed for QuesGen—namely question templates, the question quality checklist, and explicit objectives—would result in the teacher being able

to write better questions or not. The rationale for the inclusion of these features was explained in Chapters 4 and 5 which described a model for and implementation of a system designed to effect teacher professional development. The hypotheses below are based upon the expected impacts described in those chapters. In a nutshell, it was hypothesized that QuesGen would have a positive impact on the variables mentioned in these questions:

- H1:** Questions for which an educational objective has been explicitly stated using QuesGen will be more likely to address one of the stated objectives for the given unit of instruction.
- H2:** Selecting one of the question templates built into QuesGen will be associated with an increase in the number of questions targeting cognitive skills higher than the level of recall in Bloom's taxonomy.
- H3:** The use of QuesGen's question quality checklist will result in the writing of questions with fewer technical flaws.

These hypotheses were generated directly from the new functionality that is incorporated into QuesGen. They make very specific claims about what each of the new elements will do individually. To address the combined effect of the elements another hypothesis was generated:

- H4:** Questions developed using QuesGen's new features will have a higher discrimination index (DI) than questions developed without.

Each of the new functions of QuesGen could contribute to a higher DI. By having an explicitly stated objective associated with the question, it is argued that questions will be much more likely to focus on content that has been covered during instruction, and

therefore less likely to focus on extraneous content that was not necessarily covered. If extraneous content is covered in a quiz then the total expected variance in exam scores will increase, and the likelihood that the “top” students will miss questions goes up (since they are being tested on material that was not necessarily covered). Reducing the proportion of “top” students who answer correctly will lower the DI (assuming no impact on the “bottom” students). As such, aligning objectives with questions could make the “top” students more likely to answer correctly, in effect increasing DI. The use of templates will increase DI because with a greater diversity of questions that focus on higher-order thinking the proportion of students in the “bottom” group who will answer correctly from memory is likely to decrease. Third, if the question quality checklist is successful in helping instructors remove technical flaws, then the “top” students are less likely to miss questions because of grammatical errors, tricks, and the like, and the “bottom” students are less likely to answer correctly because answers are obvious, or otherwise easy to guess. The net impact of these will be to increase DI. Since it is likely to be difficult to tease out the contribution that each of these features will make to DI, H4 restricts itself to predicting a higher DI based on use of new QuesGen functionality.

7.2.2 User Satisfaction

The second set of research questions was concerned with whether or not the functionality of QuesGen had been designed in such a way that it would entice teachers to use it. The concern was that, even with QuesGen’s new features, teachers would not be persuaded to take the extra time and effort necessary to learn how to use them and, in turn, enjoy the benefits that could be gained. Before stating the research hypotheses, an informal usability study will first be described.

An informal usability study was run prior to the actual experiment to try to identify problems with the QuesGen interface. Two colleagues who were unfamiliar with this research were recruited and asked to use QuesGen to develop a 5-question quiz. Their usage of the system was observed. A number of usability issues were uncovered which fell into two categories. Most of the usability problems involved browser compatibility issues and were dealt with easily except for one. Sometimes when a user using Internet Explorer 6 would try to submit or update a question, nothing would happen when they clicked the submit button the first time. Fortunately, the intuitive reaction of the users was to click the submit button again, which allowed the button to work. This bug was not able to be worked out prior to running the study, and is believed to be caused by the 3rd party WYSIWYG HTML editor (called KHTML) that is built into QuesGen. It may have caused problems for some users and had an adverse impact on satisfaction.

The second category of issues dealt with a perceived sluggishness of the interface. This problem was also due largely to KHTML which requires running a rather large amount of javascript. The rationale for using KHTML actually came out of the second pilot study described in Chapter 3. In this study, the instructors who used QuesGen indicated that without the ability to highlight, underline, and otherwise add formatting to the text of their questions, that they didn't feel that QuesGen would be useful to them. Adding KHTML solved this problem but introduced other problems. Unfortunately, even though the Internet has been in wide use for at least ten years, there still is not a really good HTML editor that can be embedded in a web browser. KHTML seemed to be the best of the available editors at the time one was selected for inclusion in QuesGen. It

necessitated the addition of a “Page loading...” graphic that would disable the page whenever the “Add New Question” page was loaded into the browser.

With these system limitations in mind, it was unclear whether or not participants in the study would respond more or less positively to the new interface. A positive relationship is hypothesized for users of QuesGen:

H5: QuesGen users are more likely to say that they would use the system again in the future than non-QuesGen users.

The rationale for H5 is that despite the added effort required to learn the new QuesGen functionality, instructors will perceive it to provide a great added benefit, which will translate to increased satisfaction with the tool. H5 is based upon UTAUT, discussed in the last chapter, which holds that if system users perceive the system to make them better at their jobs, and not to be too much extra effort, they will be more likely to express their intention to use the system. The UTAUT model is not as simple as expressed here, but following from the rationale used to propose H5, it follows that if the perception of usefulness outweighs the perception of effort required to use it, that both satisfaction and likelihood to want to use the system will be greater than for a standard system.

7.2.3 Intervening Variables

The third set of research questions dealt with other independent variables that were likely to have an impact on question quality, namely the experience that the instructor had with teaching and writing questions, and the course content for which the questions were being written. The first set of hypotheses related to instructor experience is relatively straightforward:

- H6:** Questions written by instructors with more experience writing MCQs will have a higher DI.
- H7:** Questions written by instructors with more experience writing MCQs are more likely to be classified more highly in Bloom's taxonomy.
- H8:** Questions written by instructors with more experience writing MCQs are likely to have fewer technical flaws.
- H9:** More experienced instructors are more likely to express a desire to use QuesGen in the future.

The rationale behind H9 is perhaps not as clear as the others. The prediction is that an experienced instructor is going to be more likely to appreciate the inherent difficulty in writing MCQs than a novice instructor, and as such will be more likely to appreciate the scaffolding that QuesGen provides.

The second set of hypotheses is related to the course content. Course content is a more difficult variable to work with since it differs so much between and even with a given subject area. Not knowing ahead of time what courses would be recruited for participation in the experiment, this set of hypotheses reverts to the null hypothesis of no difference between groups:

- H10:** There will be no difference in the DI score for questions written for different course subjects.
- H11:** There will be no difference in the likelihood that a question will be categorized as a higher-order Bloom question for different course subjects.
- H12:** There will be no difference in the number of technical flaws for questions written for different course subjects.

H13: There will be no difference in satisfaction with QuesGen for different course subjects.

7.2.4 Student-Related Variables

Although there were no research questions in Chapter 1 that specifically mentioned the relationship of QuesGen to students' responses, since it was planned to collect responses from the students, it seemed in order to write some hypotheses about these responses. Hypotheses concern students' reactions to the questions with respect to the independent variables—use of QuesGen, experience of instructor, and course:

H14a: Students will perceive questions written with QuesGen to be more difficult.

H14b: Students will perceive questions written with QuesGen to be clearer.

H14c: Students will perceive questions written with QuesGen to be fairer.

H15a: Students will perceive questions written by more experienced instructors to be more difficult.

H15b: Students will perceive questions written by more experienced instructors to be clearer.

H15c: Students will perceive questions written by more experienced instructors to be fairer.

H16a: Students will perceive questions written for different courses to have different levels of difficulty.

H16b: Students will perceive questions written for different courses to have different levels of clarity.

H16c: Students will perceive questions written for different courses to have different levels of fairness.

7.2.5 Exploratory Questions

There was one exploratory research question listed in Chapter 1, RQ7 which asked if the interaction with QuesGen would have any impact on teachers' attitudes toward MCQs. Data about this question was gathered via the follow-up interviews, but no specific hypotheses were written beforehand to address this.

7.3 Conclusion

This chapter presented the hypotheses of the study. The next chapter describes the experiment that was designed and run to evaluate QuesGen.

CHAPTER 8

EXPERIMENTAL DESIGN

8.1 Introduction

First, a brief summary of the experiment is presented. QuesGen was evaluated using a controlled experiment with a single-trial, between-groups design. Two groups of instructors were used—an experimental group which will be referred to as the QuesGen group, and a control group, which will be referred to as the “standard” group. Each instructor was asked to write a set of ten MCQs about a pre-selected unit of course content. The QuesGen group used a version of the QuesGen system that had access to all of the new functionality described earlier. The standard group used a version of QuesGen where all of the new functionality had been removed so as to mirror the same type of functionality commonly available in current software. Once completed, the questions were delivered to the students of the instructors and the students were asked to complete a post-quiz survey in which they rated the questions. Next, an expert item-review panel was convened to evaluate the questions. Finally, question analysis reports (QARs) were generated and follow-up interviews were conducted with the instructors. The data gathered were used to evaluate the hypotheses described in the previous chapter.

8.2 Variables

This section will describe the variables in the experiment.

8.2.1 Independent Variables

This experiment considered three main independent variables:

- Use of the QuesGen system
- Experience of the instructor
- Course content

There were two levels of QuesGen use. The “standard” group of instructors used a very simplified version of the QuesGen system which was designed to match the functionality available in current, popular, web-based learning management systems such as Blackboard and WebCT. This system basically had a space for teachers to enter the stem of the question, and several spaces for them to enter the answer choices. Different from systems such as WebCT, the instructors were not allowed to indicate more than one correct answer. The QuesGen group had access to all of the new features described in Chapter 5, namely the dropdown menu for selecting an educational objective, question templates, the question quality checklist, and a set of video tutorials which explained what a “high quality” MCQ looks like. Except for the additional features available to the QuesGen group, the systems looked and behaved identically.

There were two experience levels. Each instructor’s level of expertise was determined during the follow-up interviews. As it turned out, about half of the participants were graduate assistants who had never written MCQs before. These instructors were categorized as the inexperienced group, and all of the other instructors were categorized as the experienced group. No attempt was made to draw any finer distinctions in levels of experience, such as number of years writing questions or the like.

There were two different courses used in the QuesGen evaluation. Both courses were 100-level General Education (GenEd) courses, meaning that all undergraduate students are required to take them regardless of major. The students are almost

exclusively freshman and sophomore level students. The first course was a kinesiology course (GKIN), which covered a range of topics related to basic health and wellness. There were approximately 24 sections of GKIN being taught by 14 different instructors at the time the experiment was run, with a total enrollment of over 700 students. The second course was called Fundamentals of Human Communication (GCOM), and comprised over 40 sections being taught by 17 different faculty members with a total enrollment of nearly 1200 students.

As such the design for this study ended up being a 2 x 2 x 2 experiment. The assignment of subjects to groups will be addressed below when the participants in the study are described more fully.

8.2.2 Dependent Variables

The dependent variables in this study have already been described in the previous chapter. The primary dependent variable was conceptualized as “question quality” and operationalized as being captured by such measures as the DI, alignment with an educational objective, adherence to best practices, and tendency to measure a higher-order cognitive skill. In addition, secondary dependent variables included instructors’ satisfaction with the system, and their attitudes toward MCQs.

8.3 Control

A number of sources of extraneous variance were anticipated to exist in this experiment. These will be discussed now, along with the measures that were taken to control them.

8.3.1 Students

Measures such as DI are sensitive to unequal variances in the population of students that respond to the questions. The student population in the experiment was controlled by carefully selecting the courses that were used for the study. As mentioned above, both courses were 100-level GenEd courses, required of all undergraduate students regardless of major. The students in the courses are almost all at the freshman or sophomore level. Very large numbers of sections are taught every semester, and given the number of sections it is not uncommon for multiple sections to be taught at the same times of day. Therefore, the population of students in any given section of a course is going to be a relatively random sampling of underclassmen. The assumption of homogeneity of variance will be tested and reported for any applicable statistics that are calculated.

8.3.2 Course Difficulty

As discussed in the last chapter, DI is also sensitive to differing levels of difficulty of content. Both of the courses were pitched at the same audience—freshman and sophomore students in the GenEd program. Both courses are taught at the 100 level and are introductory courses in their respective disciplines. These courses do not draw majors. Both courses are in the GenEd program because they are seen to cover fundamental skills and knowledge that all college graduates should have. As such, it is argued that as far as difficulty is concerned, the two courses used in the experiment are the same. While it is very difficult, if not impossible to measure the relative difficulty of these courses, there will be an analysis of the relative difficulty of the questions written by the instructors in these courses, as rated by the students.

8.3.3 Differences in the User Interface

Even though the primary independent variable was “use of QuesGen,” one primary concern was what system to use for the “standard” group. Alternatives considered included having the standard group just write their questions on paper, or having them use a pre-existing system such as Blackboard. However, the experimenter did not want the standard group to know that they were the standard group, and as such using an interface with which they were already familiar wouldn’t work. Developing a “stripped down” version of QuesGen was a relatively straightforward task. A benefit to running the study this way is that the only difference between the two groups’ interfaces was the added QuesGen functionality. This removed the variance that might have been caused if

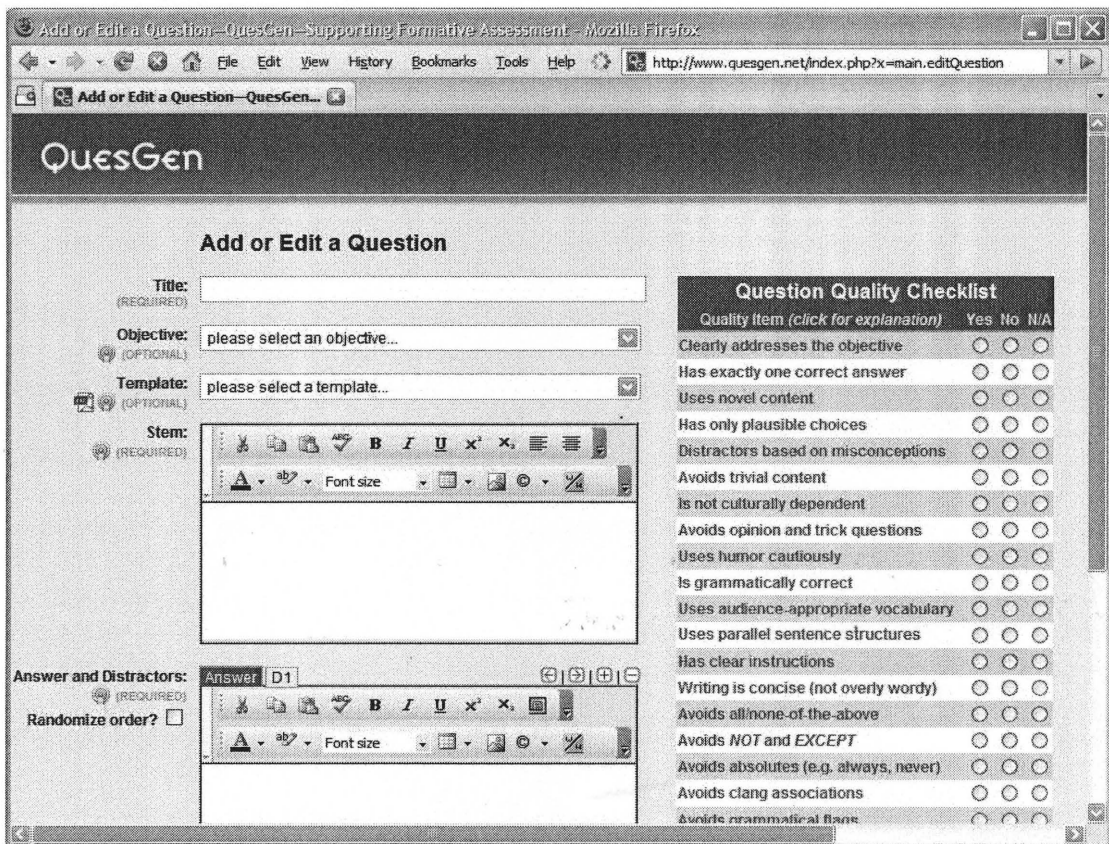


Figure 8.1 The QuesGen Interface

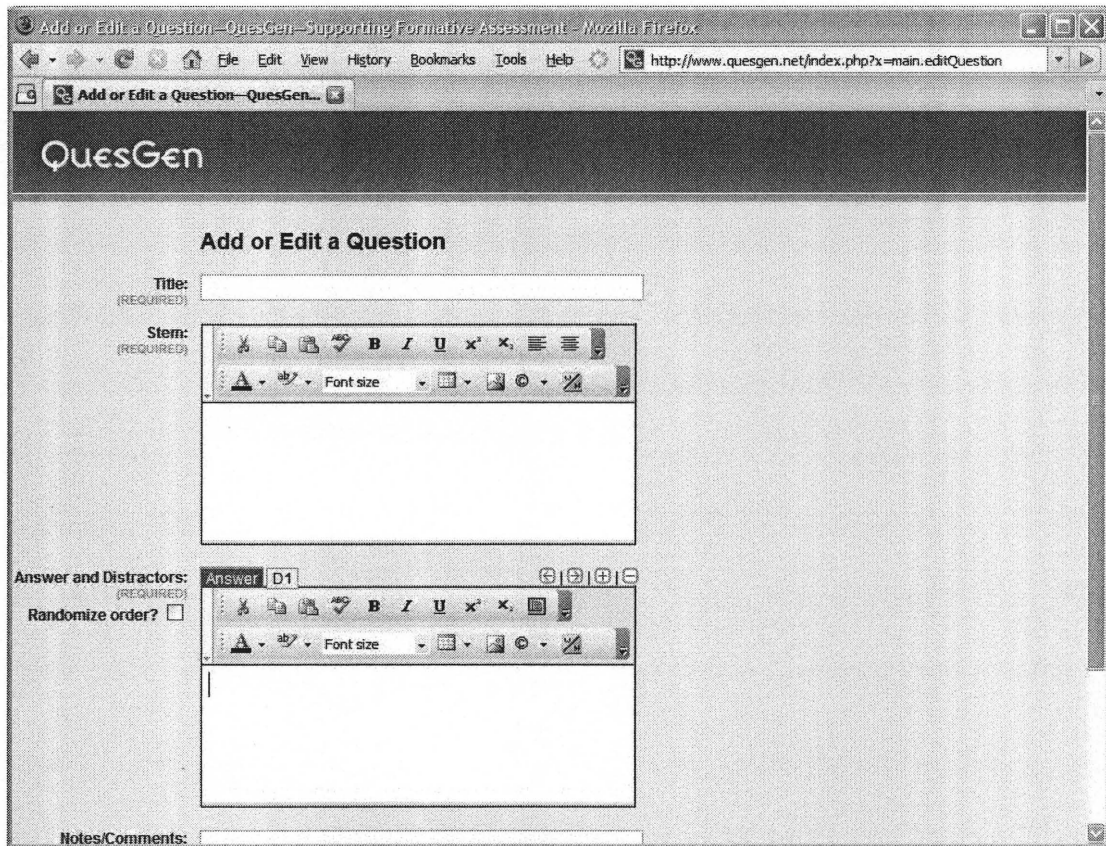


Figure 8.2 The "Standard" Interface

only one of the groups had to learn a new interface, and it also removed the possibility that the variance was due to some other feature of the system.

8.3.4 Differences in Instruction and Content

In order to control variance due to the content being considered for the quizzes, a standard video lecture was created for a single unit of instruction in each of the courses that was used. In the GKIN course this was a lecture on nutrition, and in the GCOM course this was a lecture on the uses of power in interpersonal relationships. In both cases, the content of the lecture was developed by a person who is normally responsible for the content in the course. In GKIN, the lecture was developed by the course coordinator, and in GCOM the lecture was developed by an experienced instructor who

normally teaches the course, but who happened not to be teaching it this semester. Other than asking that the objectives for the lesson be on the second slide of the lecture, no attempt was made to influence the content of the lecture. All of the instructors in the course generated questions designed to assess the content in the standard video lecture. Before they took the quizzes, the students were asked to watch the standard video lecture. Therefore within each course the variance due to differences in the delivery of course content was controlled.

8.4 Selecting Courses and Participants

The primary participants in the QuesGen study were college instructors. Given the criteria for controlling variance given above, GenEd were targeted for inclusion. Courses were chosen where there tends to be a large number of sections offered in any given semester, and in which there were at least ten different instructors teaching the course. Although the day-to-day content of the courses and the styles of their instructors may differ, each course had a single course coordinator and common exams to facilitate uniformity across sections. Support from the Dean of the General Education Program was obtained, and the dean facilitated identification of and introductions to the course coordinators of the courses that were eventually selected. An email was sent to the course coordinators approximately three months prior to the beginning of the experiment requesting an initial meeting to discuss instructor participation. Meetings were held in the late fall, prior to the beginning of the spring semester. At the initial meeting with the course coordinator, the nature and purpose of the QuesGen study was explained, along with the potential benefit to the course for instructors that participated. Coordinators

were asked to facilitate the recruitment of the individual instructors, to help identify content for the unit of instruction to be used in the study, and to help choose a time in the semester at which it would be appropriate to run the study. The potential pool of participants was seventeen instructors in the GCOM course and thirteen instructors in the GKIN course.

Approximately one month prior to the target start date of the study, an email was sent to the target instructors. All instructors were on the faculty of a medium-sized, rural, mid-Atlantic, teaching university. In the GKIN course, the coordinator was able to give a strong assurance that most instructors would participate in the study because most of the instructors were graduate assistants and their instruction was closely guided by the lead instructors in the course. The lead instructors were consulted by the coordinator prior to their agreement to participate in the study. In the GCOM course the coordinator had much looser control over the instructors, and so in addition to the introductory email, the investigator went to a meeting of the instructors and made an in-person pitch.

When recruiting, the nature of the experiment was explained in full to the course coordinator, but that person was asked not to share the details with the instructors who would participate in the study. When the study was introduced to the instructors it was explained that QuesGen was a system designed to help teachers write better multiple choice questions, but no further explanation as to the rationale or mechanism was given. After the instructors in the courses agreed to participate, they were each sent an email with the start date, list of tasks and schedule of their participation. They were also asked to provide course rosters for all of the sections they were teaching that semester. Explicit consent was not obtained until instructors logged into the QuesGen site for the first time.

8.5 Experimental Procedure

This section gives a narrative account of how the QuesGen study was planned.

8.5.1 Week 1: Instructors Write Questions

On the morning of day one of the study, an email was sent to all of the participating instructors with detailed instructions for participation (see Appendix G). The study was timed to start on a Monday morning, and instructors were asked to have logged into QuesGen, watched the tutorial and content videos, and written their ten questions by noon on Thursday. In actuality, since students were not to begin taking the quizzes until the following Monday, it was only necessary that instructors have completed their questions by Sunday evening, day seven of the study. However, knowing that instructors are busy people a due date much earlier than was necessary was chosen. As expected, a number of the instructors logged in immediately and had completed their questions by the requested due date and time. Knowing that these instructors were done freed the experimenter to focus on motivating the less prompt instructors to participate.

Emails were sent daily to remind instructors of the study. These emails were phrased as announcements, updates, or status checks and were written with the intention of keeping the QuesGen study in the mind of the instructors, without coming across as too “pushy” or insistent about participation since it was not desired to be annoying or otherwise antagonistic towards the instructors.

Some technical issues arose from time to time, such as instructors not remembering their passwords, or having lost or deleted their original instructional email. All emails or phone calls from instructors were returned as soon as possible after they arrived, usually within minutes, and in no case more than a couple of hours.

Also during this week a 1-page flyer announcing the study to students was delivered to each of the instructors. Enough copies of the flyer were printed so that each student enrolled in the instructors' classes could have a copy. The flyers were delivered to instructors' departmental mailboxes. The purpose of the flyer was several-fold. First, the flyers were a tangible way to remind the teachers about the study. Second, it was thought that some college instructors' reluctance to waste paper might convince them to pass out the flyers, even if they had not been strongly committed to participating in the study in the first place. Third, it was hoped that if instructors distributed the flyers to students that their motivation to complete questions would increase, since now they would need to follow through with an activity that had been announced to the students. Fourth, and perhaps least importantly, the flyer was designed to prime the students' awareness of the study and prepare them for the email they would receive the following week asking for their participation.

Upon logging into the QuesGen system, instructors were first presented with an online consent form asking them for their consent to collect data for use in the study. Upon clicking the "I consent" button a PDF version of the consent form with the instructor's name was emailed to them and also to the experimenter. The instructor was then taken to the QuesGen home page which contained instructions for participation. First, the instructor was asked to watch the video lecture about which the instructor would later be asked to write questions. The video lectures were created by one of the instructors of the course who was not participating in the study, and based on the content from the standard syllabus of the course. After watching this video, the instructor was asked to watch a video tutorial showing how to use QuesGen's question entering

interface. There were two different video tutorials—one explaining the QuesGen interface, and one explaining the standard interface. The QuesGen interface also had links to several other video tutorials which discussed four topics: educational objectives, writing good questions, writing good distractors, and question templates. It was possible for instructors using either interface to write questions without watching any videos at all.

After watching the videos, the instructors were asked to write their questions. Depending upon the group to which each had been assigned, each was presented either with the QuesGen interface or the standard interface. Upon completing their ten questions, each of the instructors was asked to log back in to QuesGen and complete the user satisfaction study. They were asked to do so as soon as possible after they had written their questions so that their impressions of the system would still be fresh in their minds.

More detailed results of this week and the following weeks are contained in the next chapter. The next section describes week two of the study in which students took the quizzes.

8.5.2 Week 2: Students Take Quizzes and Surveys

On the morning of the second Monday of the study, the first day of week two, an email (see Appendix H) was sent to all of the students in all of the sections being taught by teachers who had completed their ten questions during week one. The emails were personalized and provided each student with a username and password as well as the URL of the website where they should go to log in and participate in the study.

Two methods were used to motivate the students to take the quiz. First, both the flyer and the introductory email let the students know of their chance to win one of three

\$25 gift certificates to be awarded at random to students who completed both the quiz and the survey. Second, the participating instructors in the study were asked to assign a small grade or extra credit for participation in the quiz. The goal of the grade was to provide enough motivation for students to participate, but not so much motivation that they would be tempted to cheat. A similar motivation strategy was adopted successfully in the first pilot study (see Chapter 3).

Students were asked to complete their quizzes prior to their second class meeting of the week. Some classes met Monday-Wednesday-Friday, and others met Tuesday-Thursday. It turned out that the week of the study was the week immediately prior to spring break. There was not really a hard and fast deadline for student participation as long as sufficient numbers had participated prior to the end of the week's study. All students viewed the same interface regardless of the interface with which the questions were written. As in the previous week, the experimenter remained on call to respond to any technical difficulties that might arise over the course of the week.

8.5.3 Week 3: Expert Panel Reviews Questions

A panel was recruited to perform an expert review of all of the items that had been written by the instructors. The two judges were graduate students in the school of assessment and measurement at the university where the study took place. Both judges had completed coursework on the theory of assessment design. One of the judges worked professionally as an item review consultant, and the other, who already had a PhD in plant pathology, was in charge of program assessment for the department where she was a tenured faculty member. The judges knew each other and had worked together before in their assessment and measurement graduate program.

The judges and the experimenter met for the entire day on Tuesday during the third week of the study. The first thing the judges did was to watch the content video for one of the participating courses. They were given a printed copy of the slides which contained the study objectives. They then were asked collaboratively to rate several of the questions that were written by the participating instructors. The judges were instructed to go through each of the items on the item-review instrument and discuss it until they had established a common understanding of the meaning of the items. Then they each rated five questions on their own and then agreement was calculated. Once the judges were satisfied that they both were interpreting the instrument in the same way, they went about rating all of the rest of the items for the first course. The pool of questions for the first course consisted of 100 questions. The order of the questions had been randomized so as to mix up questions that had been generated with the QuesGen and standard interfaces. An online system was built for them to do the ratings, and both judges received the questions in the same order. The judges took regular breaks since rating the questions was somewhat tedious.

By the end of the first day, the judges were only able to complete the questions from the first course. Since the question rating system was online, the instructors were allowed to complete the ratings on their own time. It was agreed that they would complete all ratings by the following Friday. One of the judges completed all ratings by the scheduled day. The other judge, unfortunately, had a death in the family later in the week and required an additional two weeks before she could complete all of the ratings. Despite the time lag, overall agreement between the judges was 84% over a total of 5780 ratings. Analysis of these results will be reported in the next chapter.

8.5.4 Week 4: Follow-up Interviews with Instructors

As described in the last chapter, follow-up interviews were conducted with the instructors. Although follow-up interviews had originally been planned for week four of the study, since the judging was not completed until almost the end of the fifth week, interviewing didn't begin until week six, which meant that five weeks had elapsed between the time that most instructors had completed writing their questions and the time that the interviews took place.

Interview times were scheduled individually with the instructors via email and the experimenter met the instructors in their offices. In several cases the instructors had shared offices and the interviews were moved to adjacent rooms, or other unoccupied areas. The experimenter asked permission to video tape the interviews, which was granted in all but one case. After signed consent was obtained, the video camera was started and the interview began. It was explained that the purpose of the tape was to free the interviewer from taking too many notes and allow him to focus on the questions to be asked. The interviews proceeded from there.

The interviews lasted about one hour each and were broken up roughly into two parts. In part one, following the interview guide in Appendix E, the interviewer obtained information about the instructor's field of expertise and teaching experience. Following that a significant amount of time was generally spent discussing the character of the students at the university, and the teacher's perception of his or her role with respect to the students. The next segment of the interview typically focused on either the course which the instructor was teaching for which she or he had written questions, and perceptions of that course's importance and place within the students' overall course of

study. Finally the interviewer probed the instructor's typical practices with respect to testing and assessment, and their attitudes to various forms of assessment including multiple choice and short answer questions.

In part two of the interview, the interviewer produced a Question Analysis Report (QAR—see Appendix D) which contained the summarized results of the evaluations of the questions this instructor had written. In most cases it was necessary to step through and explain each part of the QAR to the instructor in detail so that he or she would be able to use it to interpret the quality of his or her own questions. This took considerable time. The interview concluded by getting the instructor's reaction to the evaluations seen on the report.

8.6 Data Analysis

In total, twenty-one instructors, representing 1236 students wrote a total of 210 questions that were delivered using QuesGen. Over 800 students took the quizzes and over 600 responded to the follow-up surveys. Those questions were evaluated by two judges who together completed 11,560 ratings of question quality. Follow-up interviews were conducted with the participating instructors. The description of the analysis of all of this data is described in the next chapter.

CHAPTER 9

RESULTS

9.1 Introduction

Chapter 7 described 22 hypotheses. The table below gives a brief synopsis of the findings with respect to each of the hypotheses. The rest of this chapter will describe the results of the QuesGen study in detail.

Hypothesis	
Supported?	Statistical test results and significance
H1:	Questions for which an educational objective has been explicitly stated using QuesGen will be more likely to address one of the stated objectives for the given unit of instruction.
Not Supported:	$\chi^2 = 0.0271, df = 1, p = 0.8692, N = 200$
H2:	Selecting one of the question templates built into QuesGen will be associated with an increase in the number of questions targeting cognitive skills higher than the level of recall in Bloom's taxonomy.
Supported:	$\chi^2 = 12.6211, df = 1, p = 0.0004, N = 200$
H3:	The use of QuesGen's question quality checklist will result in the writing of questions with fewer technical flaws.
Not Supported:	$U = 8794, 1\text{-sided } p = 0.1286, N = 200$
H4:	Questions developed using QuesGen's new features will have a higher discrimination index (DI) than questions developed without.
Not Supported:	$DI_{qg} = 0.41 > DI_{std} = 0.39, U = 10,375.5, p = 0.2134, N = 200$
H5:	QuesGen users are more likely to say that they would use the system again in the future than non-QuesGen users.
Not Supported:	$BI_{qg} = 10.7 < BI_{std} = 11.0, U = 53.5, p = 0.5, N = 17$
H6:	Questions written by instructors with more experience writing MCQs will have a higher DI.
Reversed:	$DI_{xp} = 0.36 < DI_{noxp} = 0.46, U = 9172, p = 0.0024, N = 200$
H7:	Questions written by instructors with more experience writing MCQs are more likely to be classified more highly in Bloom's taxonomy.
Supported:	$\chi^2 = 51.197, df = 1, p < 0.0001, N = 200$
H8:	Questions written by instructors with more experience writing MCQs are likely to have fewer technical flaws.
Supported:	$U = 27,264.5, p < 0.0001, N = 200$
H9:	More experienced instructors are more likely to express a desire to use QuesGen in the future.
Supported:	$BI_{xp} = 13.4 > BI_{noxp} = 6.3, U = 30.0, p = 0.0088, N = 17$

Hypothesis	
Supported?	Statistical test results and significance
H10: There will be no difference in the DI score for questions written for different course subjects.	
Not Supported:	$DI_{GKIN} = 0.43 > DI_{GCOM} = 0.37$, $U = 9409$, $p = 0.0586$, $N=200$
H11: There will be no difference in the likelihood that a question will be categorized as a higher-order Bloom question for different course subjects.	
Not Supported:	$GCOM > GKIN$, $\chi^2 = 67.124$, $df = 1$, $p < 0.0001$, $N=200$
H12: There will be no difference in the number of technical flaws for questions written for different course subjects.	
Not Supported:	$GCOM > GKIN$ $U = 44,937.5$, $p < 0.0001$, $N=200$
H13: There will be no difference in satisfaction with QuesGen for different course subjects.	
Fail to Reject:	$QUIS_{GCOM} = 35.78$, $QUIS_{GKIN} = 35.33$, $U = 87.5$, $p = 0.4473$, $N=18$
H14a: Students will perceive questions written with QuesGen to be more difficult.	
Reversed:	$SD_{qg} = 2.17 < SD_{std} = 2.31$, $U = 21,114$, $p = 0.0027$, $N=200$
H14b: Students will perceive questions written with QuesGen to be clearer.	
Supported:	$SC_{qg} = 4.31 > SC_{std} = 4.26$, $U = 26,038$, $p = 0.0310$, $N=200$
H14c: Students will perceive questions written with QuesGen to be fairer.	
Weak Support:	$SF_{qg} = 4.17 > SF_{std} = 4.13$ $U = 25,426$, $p = 0.0987$, $N=200$
H15a: Students will perceive questions written by more experienced instructors to be more difficult.	
Reversed:	$SD_{xp} = 2.14 < SD_{noxp} = 2.46$, $U = 39,996$, $p < 0.0001$, $N=200$
H15b: Students will perceive questions written by more experienced instructors to be clearer.	
Supported:	$SC_{xp} = 4.34 > SC_{noxp} = 4.17$, $U = 26,158$, $p < 0.0001$, $N=200$
H15c: Students will perceive questions written by more experienced instructors to be fairer.	
Supported:	$SF_{xp} = 4.24 > SF_{noxp} = 3.98$ $U = 24,090$, $p < 0.0001$, $N=200$
H16a: Students will perceive questions written for different courses to have different levels of difficulty.	
Supported:	$SD_{GCOM} = 2.13 < SD_{GKIN} = 2.40$, $U = 33,126$, $p < 0.0001$, $N=200$
H16b: Students will perceive questions written for different courses to have different levels of clarity.	
Not Supported:	$SC_{GCOM} = 4.28$, $SC_{GKIN} = 4.26$, $U = 41,170$, $p = 0.3549$, $N=200$
H16c: Students will perceive questions written for different courses to have different levels of fairness.	
Supported:	$SF_{GCOM} = 4.21 > SF_{GKIN} = 4.06$ $U = 44,958$, $p < 0.0001$; $N=200$
<p>U = Mann-Whitney U, 1-sided Z calculated using normal approximation SD = Student perceptions of question difficulty, scale 1-5, 5=more difficult SC = Student perceptions of question clarity, 5=more clear SF = Student perceptions of question fairness, 5=more fair</p>	

Table 9.1 Synopsis of QuesGen Study Results

9.2 Experiment Overview

On day one of the study an introductory email (Appendix G) was sent out to thirty instructors: 17 from a general education communications course (GCOM), and 13 from a general education kinesiology course (GKIN). Prior to sending the email the instructors had been randomly assigned to either the test condition (QuesGen interface) or the control condition (standard interface). By the end of week one of the study twenty-one instructors had successfully completed ten MCQs. The breakdown of instructors by condition was as follows:

Interface	Course	
	GCOM	GKIN
QuesGen Interface	6	5
Standard Interface	5	5

Table 9.2 Breakdown of QuesGen Participating Instructors

Upon inspection of the questions generated by the instructors, it was found that one of the GCOM-QuesGen instructors had extreme difficulties using the system, to the point that those questions needed to be discarded. Figure 9.1 shows an example of one of this instructor's poorly formatted questions. Most of the questions had formatting issues like this one which left students unable to select a correct answer, even if they knew what it was. This left five instructors in each cell of the table above. After completing their questions, the instructors were asked to log into the system again and complete the follow-up survey to determine their satisfaction with the system. Eighteen of the twenty-one instructors completed this follow-up survey.

Question 3 of 10

You have just learned that your minister is having an affair with one of his congregants. You are about to ask your minister for a loan. Which type of power resource might you most successfully use?

A. Information

Expertise

Knowledge

B. Physical

Figure 9.1 Question Exhibiting Formatting Problems
Instructor's data was removed from the study—all answer options
are given as the first choice

On day seven of the study, one week after the instructors received their email, an email (Appendix H) was sent to all of the students enrolled in the courses taught by the twenty-one instructors who had completed questions. This email was sent to 1236 students. By the end of the twelfth day of the study 820 students had completed their online quizzes and 636 had completed the follow-up survey. Although relatively few students reported having trouble either logging in or using the system, an uncaught bug in the system prevented the quiz and survey submissions of an unknown number of students from being recorded. Students were not required to take the quiz in all courses—some instructors counted it as a minor assignment, while others made participation worth extra credit. At least ten students per participating instructor responded to the quiz and survey.

On day fifteen of the study, two expert judges spent an entire day rating the 200 questions that had been written by the instructors. The judges spent the morning training and calibrating their responses and the afternoon rating the first 100 questions. Given the

number of questions and their ability to complete the rating online, the judges were allowed to complete their ratings on their own time. One judge completed the ratings within three days of the first session. Due to a death in the family, the second judge did not complete rating the questions until approximately two weeks later. The judges agreed 84% of the time over a total of 4830 ratings (rating 23 items on each of 210 questions). Cohen's Kappa was calculated to gauge inter-rater reliability, but this produced a very low score. It was determined that the low Kappa value was due to the low variance in the response patterns of the judges. In general, the questions written by the instructors were deemed to violate very few of the best practices against which they were being judged. Viera et al. (2005), explain that in such cases Kappa is not a very good indicator of inter-rater reliability, and they indicate that in such cases the raw agreement percentage may be used.

9.3 Limitations of the Study Design

Several factors limited the ability of this study to determine QuesGen's effectiveness at getting teachers to write better multiple-choice questions. Three limitations will be discussed here: floor effects in the item-review instrument, the degree to which instructors actually used the new functionality, and a confound between the experience level and courses that instructors were teaching.

The twenty-one elements of question quality that were addressed in the item-review instrument (Appendix C) were taken directly from the quantitative empirical literature on MCQ construction. As such, it was expected that high quality questions would conform to these best practices to a greater extent than low quality questions.

What was harder to predict was that few questions individually would express a great number of flaws. In other words, while many questions had flaws, and while the flaws they exhibited varied across the possible categories, it was unusual for any single question to have more than two or three flaws out of the possible total of twenty-one. Furthermore, the literature did not provide enough information to generate a meaningful weighting system, whereby certain flaws might be quantitatively expressed as more serious than other flaws. Therefore, even though the judges were in high agreement over which flaws were contained in the items, the overall scores generated by the item-review instrument had fairly low variance, and as a result provided poor resolution of quality between questions.

Fortunately, despite the low resolution, useful data was able to be gleaned from the item-review instrument. While more experience using the item-review instrument could have helped to correct for the resolution issues, the results of this study may suggest that technical flaws are not the place to focus the energy of a tool like QuesGen, and that the problems with questions may come from other sectors. These issues will be addressed in more depth in the discussion chapter to follow this one.

The second limitation of this study was the degree to which participating instructors actually used the new functionality embedded within QuesGen. There were ten instructors in the QuesGen group who had access to the new features, but of these ten, only six used any of the features, and only three used the features to any great extent. Because of these usage patterns it is difficult to say whether or not QuesGen would have had a bigger impact if more people had used it. The few instructors who used the new features were enthusiastic about them, but it is difficult to say whether their enthusiasm is

due to the features—it is possible that these people have an “early adopter” personality type which gets excited about any new features regardless of their effectiveness. Some of the results below offer insight into what might be the features to focus on for the next round of QuesGen development and evaluation. There may be other ways to construct the study, such as greater payment, or face to face training, so that there will be a more uniform adoption of the features by the participants.

A third major limitation of the study is a coincidental overlap between the courses being taught and the experience levels of the participating instructors. Almost all of the instructors in the GKin course were graduate assistants (GA's) and this was their first time ever writing MCQs. Conversely all of the GCOM faculty members were experienced question writers, in that they had all been writing MCQs several times a semester for at least two semesters, but in most cases many more.

Question Writing Experience	Course	
	GCOM	GKin
None	0	8
At least 1 semester	10	2

Table 9.3 Distribution of Instructors by Course and Question Writing Experience

On one hand, the stark contrast between beginners and more experienced teachers was beneficial in that the data seemed to show some real differences that are more likely to be explained by the experience gap than by the differences in course material. On the other hand, separating the effects of course material from experience is impossible from a statistical standpoint, making stronger conclusions from the analyses of the data collected not viable. Future studies will need to do a better job of controlling the aspects of the participating instructors.

A fourth limitation of the study is its duration. The study collected instructor questions at a single point in time. Unless the mechanisms built into QuesGen are powerful enough to cause a nearly instantaneous improvement in question quality, the results of the study are unlikely to reveal a significant effect on question quality due to use of the system. Since the goal of QuesGen is to teach instructors how to write better questions, and since that type of learning is something that happens over time, it is reasonable to think that a longitudinal study may be needed to uncover the degree to which QuesGen is able to have an impact on MCQ-writing ability.

Finally, it is important to acknowledge that the QuesGen system itself was not perfect. Some of the noise in the data is undoubtedly due to glitches with the interface. This is a problem in the development of any type of web-based application that is designed to be deployed to a diverse audience. It is impossible to predict all of the idiosyncratic differences between web-based system users' computer configurations. While QuesGen was good enough to be used to gather data for a study such as this one, it is not a system that is ready for "prime time" as of yet. Indeed, a study such as this one was necessary to help seed ideas for and focus the development of the next version of the system. Given these limitations, it is now time to take a look at what the study found.

9.4 Impact of QuesGen Functionality

This section will discuss the impact of QuesGen's new functionality. The first set of hypotheses (H1-H4) all addressed whether or not the new features that were built into QuesGen would have an impact on question quality. In this section the data pertaining to these hypotheses will be presented and analyzed.

9.4.1 Aligning Objectives with Questions

The first hypothesis (H1) was that if instructors used QuesGen to select an objective from a dropdown list, that the judges would be more likely to agree that the question addressed a stated objective from the lecture (Appendix C, Section 1). Regardless of what interface the instructor used to write the question, the judges were presented with the list of objectives for that lecture and asked to indicate to what extent they found the question addressed each of the objectives. Therefore the degree to which an objective was addressed was rated for all 200 questions. Since both the independent variable (objective selected or not) and the dependent variable (whether or not judges said the question addressed a stated objective) were categorical, a chi-squared test was run. The chi-squared analysis did not support the hypothesis. In fact, instructors who used QuesGen to select an objective were slightly, but not significantly, *less* likely to have their question judged as addressing one of the objectives of the lesson.

QuesGen Usage	Judge's Evaluation		Total
	No Objective Addressed	Objective Addressed	
No Objective Selected	41 26.8%	112 73.2%	153 76.5%
Objective Selected	13 27.66%	34 72.34%	47 23.5%
Total	54 27%	146 73%	200 100%
Chi-square = 0.0271, p = 0.8692			

Table 9.4 Objectives Selected with QuesGen vs. Judge's Evaluation that Objectives were Addressed by the Question

In total, 200 questions were written by the instructors. Half of those questions were written by instructors using the QuesGen interface. Of those, 47 questions had an objective explicitly associated with them via the dropdown menu supplied by QuesGen, making this the “most popular” of QuesGen's features. Despite having explicitly

selected an objective with which to associate a question, the proportion of questions for which the judges agreed that the question addressed one of the stated objectives was the same as for the group of instructors who did not have the dropdown list. Chi-squared analyses were run which controlled for instructor experience, and the course being taught. None of the tests yielded significant results.

QuesGen Usage	Control Variable	Judges' Evaluation that an Objective was Addressed or Not
No Objective Selected	Experience	$\chi^2 = 2.37, p = 0.1237, N = 153$
Objective Selected	Experience	$\chi^2 = 0.33, p = 0.5663, N = 47$
No Objective Selected	Course	$\chi^2 = 0.02, p = 0.8791, N = 153$
Objective Selected	Course	$\chi^2 = 1.39, p = 0.2381, N = 47$

Table 9.5 Results of Chi-squared Tests of Adherence to Objectives Controlling for Instructor Experience and Course Taught

9.4.2 Using Question Templates

H2 predicted that questions for which an instructor had specified a question template would be more likely to address higher-order thinking as rated by the expert judges. This feature of QuesGen was not used a great deal—a total of 13 questions out of 200 had templates associated with them explicitly. Of those, ten questions with templates were written by a single instructor, two by a second, and one by a third. The statistical analysis below is based only upon these thirteen questions. The data reported in 0 support the hypothesis that using QuesGen's templates is associated with higher-order questions. The small number of questions for which instructors chose templates (thirteen), and the fact that the majority of the questions with templates were generated by a single instructor argues that caution should be taken not to overemphasize the significance of these results. On the other hand, the fact that a significant result was obtained with such a small sample size with a non-parametric test indicates that the size of the effect caused by the question templates may be relatively large.

QuesGen Usage	Judge's Evaluation		Total
	Recall	Higher than Recall	
No Template Selected	122.5 65.51%	64.5 34.49%	187 93.5%
Template Selected	4 30.77%	9 69.23%	13 6.5%
Total	126.5 63.25%	73.5 36.75%	200 100%
Chi-square = 12.6211, p = 0.0004			

Table 9.6 Template Selected with QuesGen vs. Judges Evaluation that Questions Addressed Cognitive Skills Higher than Recall

The follow-up interviews with instructors provided more information on the use of the templates. In speaking with the instructor who used a template for all ten of her questions, she was very enthusiastic about the template feature of QuesGen. She said that she had downloaded the PDF version of the templates and had used it several times since participating in the study to write MCQs for her classroom assessments. The instructor who had specified templates for only two of his questions reported that despite only having explicitly associated two with questions, he had, in fact, used the templates continuously throughout the process, and found them very useful. It was realized that the system usage logs were not capturing clicks on content such as the pop-ups that explained how to use the templates. As such it is possible that the actual use of the templates was higher than is indicated in the above table, although it is unclear how this additional usage would impact the statistics. The initial analysis seems to indicate that additional effort to encourage template use is warranted.

Another important statistic to note is that without templates, 65.5% of the questions written were deemed to address recall-level cognitive skills. That the majority of the questions asked in these college level courses did not address higher-order thinking

lends support to the basic premise of this research, that instructors do not write very high quality questions.

9.4.3 Using the Question Quality Checklist

H3 predicted that QuesGen users who used the question quality checklist would write questions that the judges found to have fewer flaws vis-à-vis compliance with the best practices of item writing as expressed in items 3 through 23 on the item-review instrument (Appendix C). The expert judges rated each question against each of the items on the item-review instrument and indicated if they found the question to be in compliance, out of compliance, or somewhere in between. These values were scored 2, 0, and 1 respectively. A score was calculated for each question by adding the points accumulated for all checklist items. That meant that each question could score between zero and forty-two, two times the number of items on the item-review instrument. Shapiro-Wilk tests for normality were run for both the checklist and non-checklist groups of questions, and found that in both cases data were not normally distributed ($W_{\text{checklist}}=0.94, p=0.03, W_{\text{non-checklist}}=0.91, p<0.0001$). As such the non-parametric Mann-Whitney U test was selected to compare the values for the two groups. While the average score for compliance with the best practices was slightly higher for the people who used the checklist, and for users of QuesGen overall, no significant difference was found between those items where instructors had used the checklist and those where the instructor had not ($U=8794, 1\text{-sided } p=0.1286, N=200$). Since it was plausible that instructors using the QuesGen interface looked at the checklist, even if they didn't actually check off any boxes, U was calculated to see if QuesGen users were in higher compliance than standard users. No significant difference in compliance with best

practices was found between items written with the QuesGen interface as opposed to the standard interface ($U=24,783.5$, 1-sided $p=0.2442$, $N=200$).

The follow-up interviews that were conducted with the instructors provided more depth and insight into these results. One of the QuesGen users, who made an effort to complete the checklist for all ten of her questions, made the comment that although she liked the checklist at first, after the third or fourth time through, she felt like she had internalized the items on the checklist and didn't really need to look at them anymore. She admitted that by the end of her ten questions she was just checking down the list without spending time actually thinking about whether or not she had written a question in compliance with that item or not. Another instructor indicated that he clicked on the pop-up help to learn about the meaning of some of the checklist items, and that although he didn't check off items on the checklist explicitly, he did look at it and consider the checklist items when writing his questions. This is another situation, as with the question templates, in which instructors "used" QuesGen functionality but the system logs were not able to capture that usage.

How does QuesGen's question quality checklist contribute to question quality? The statistical analysis presented above indicates that it contributes little, if at all. Analyses failed to find any significant relationships between use of the checklist and judges' likelihood to rate a question as having fewer technical flaws. Conversations with instructors who used the checklist indicated that while initially useful, the salience of the checklist soon waned as instructors quickly internalized the items on the list. While intuitively it seems like it would be a mistake to remove the checklist altogether because of the learning about high-quality questions it potentially brings to instructors, the results

of this test suggest that the form of the checklist should be significantly altered in future versions of QuesGen.

9.5 QuesGen and the Discrimination Index (DI)

What did the DI data have to say about question quality? 803 students took quizzes as part of the QuesGen study. DI was calculated for 200 questions using an average of just over 40 students per instructor (mean=40.15, min=9, max=100, $\sigma=22.6$). This section will address the three hypotheses (H4, H6 and H10) that made predictions about DI.

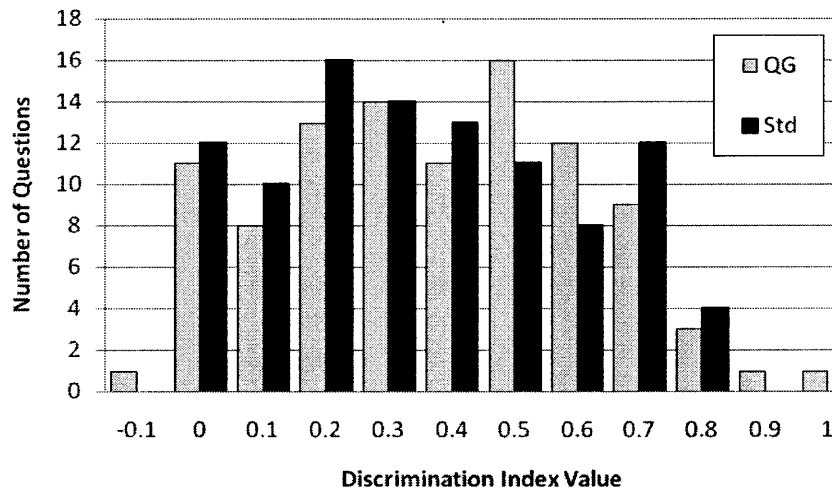


Figure 9.2 Distribution of DI Scores for Different System User Groups

H4 predicted that questions developed using QuesGen's new features will have higher DI than questions developed without. A Shapiro-Wilk statistic found that the DI values significantly departed from the normal distribution ($W=0.98$, $p=0.0028$). Therefore a Mann-Whitney U statistic was calculated to determine whether there was a difference between the DI scores of instructors in the two groups. No significant differences were found between the two groups ($U=10,375.5$, $p=0.2134$, $N=200$). Figure

9.2 shows the distribution of scores in the two groups. The average DI score for the QuesGen group was slightly higher than the standard group—0.41 versus 0.39. As such, H4 was not supported.

H6 predicted that questions written by instructors with more experience writing MCQs would have higher DI scores. The opposite turned out to be the case. Instructors who had never written MCQs before wrote questions with an average DI of 0.46, which was significantly higher than the average for more experienced instructors, whose DI was .36 ($U=9172, p=0.0024, N=200$).

H10 predicted that there would be no difference in the DI scores for questions written for different course subjects. Given the result in H6 and the previously cited overlap between the course and experience variables in this study, it is not surprising that the GKIN instructors had a nearly significantly higher DI than GCOM instructors ($DI_{GKIN}=0.43 > DI_{GCOM}=0.37, U=9409, p=0.0586, N=200$). H10 was not supported. However, the fact that the relationship between DI scores for H6 is more significant than for H10 makes it at least plausible that the relationship observed between the courses is explained mostly by the instructors' levels of experience.

The DI results were compared to the expert judges' likelihood of rating a question as addressing higher-order Bloom's taxonomy levels. The means for the recall group versus the higher-than-recall group were extraordinarily close together at 0.400182 and 0.400127 respectively. Although not valid, since the data violate the assumption of normality, it is interesting to report that the F statistic was 0.0000, $p=0.9983$. Non-parametric tests confirmed that no significant relationship exists between these variables ($U=29,633, 2\text{-sided } p=0.8861$). These results suggest that DI and Bloom level address

very different aspects of question quality. While the differences between lower and higher order cognitive skills as classified by Bloom are relatively straightforward to grasp, the meaning of lower and higher DI scores is still not apparent.

In summary, hypotheses 4 and 10 were not supported, and H6 showed a significant relationship in the opposite direction from the one predicted. A great deal of time is spent in the next chapter addressing the problem of how to interpret these DI scores.

9.6 Behavioral Intention to Use QuesGen

H5 predicted that instructors using the QuesGen interface would be more likely than non-QuesGen users to say that they would use the system again in the future. This hypothesis was made with the reasoning that users would perceive QuesGen's added functionality as a benefit worth coming back for. The reason that users' stated intention is a variable of interest is that research has found that the best prediction for whether or not users will use a system is what they say that they are likely to do. This is operationalized in the IS literature as the behavioral intention to use a system, and it is measured by the UTAUT instrument (Venkatesh et al, 2003). Data for both the QuesGen and non-QuesGen groups did not differ significantly from the normal distribution ($W_{QG}=0.88$ $p=0.31$, $W_{noQG}=0.9$ $p=0.20$). No significant differences were found between QuesGen and non-QuesGen users on this variable ($t=0.40$, $p=0.69$, $N=16$). Therefore H5 was not supported.

H9 predicted that more experienced users would be more likely to indicate that they would use QuesGen again in the future if it were available. This hypothesis was made reasoning that more experienced instructors would be more likely to be able to spot

the utility embedded within the tool. The data gathered for these two groups were normally distributed ($W_{xp}=0.90$ $p=0.39$, $W_{noxp}=0.94$ $p=0.52$). A t -test indicated that the average BI value for experienced teachers was significantly higher ($BI_{xp}=14.4$, $BI_{noxp}=6.33$, $t=-4.14$, $p=0.001$). Given the small sample size ($N=16$), a power analysis was calculated and indicated that with $\alpha = 0.05$, there was a 98.9% chance of detecting a difference if one actually existed. For $\alpha = 0.01$, the chance of detecting a difference decreased to 91.9%. H_9 was supported by these results.

Although no hypotheses were made with respect to the impact of course on behavioral intention to use QuesGen, given the overlap between the experience and course groups, differences were calculated for these groups. The data met the normality condition ($W_{GCOM}=0.91$ $p=0.31$, $W_{GKIN}=0.93$ $p=0.57$), and a t -test was performed. As expected given the results of the previous test, the GCOM instructors were found to be much more likely to express their desire to use QuesGen in the future ($BI_{GCOM}=15.2$, $BI_{GKIN}=6.42$, $t=-5.53$, $p<0.0001$). For $\alpha = 0.01$, the chance of detecting a difference was 99.4%. For $\alpha = 0.001$, the chance of detecting a difference decreased to 91.4%.

Finally, given the apparent strong effect of course on BI, an ANOVA was run in an attempt to determine the relative degree to which experience and course each accounted for the BI scores. 0 and 0 display the results of the ANOVA. The model is very significantly related to the differences in BI scores observed, and has an overall R^2 value of 0.687. When the independent variables are compared, course appears to account for a much more significant degree of the variance than does instructor experience.

Source	DF	SS	Mean Square	F	Pr > F
Model	2	304.861	152.43	14.27	0.0005
Error	13	138.889	10.684		
Total	15	443.75			

Table 9.7 ANOVA Results for Impact of Course and Instructor Experience on Behavioral Intention to Use QuesGen

Source	DF	Type I SS	Mean Square	F	Pr > F
Course	1	304.48	304.48	28.50	0.0001
Experience	1	0.38	0.38	0.04	0.8531
Source	DF	Type III SS	Mean Square	F	Pr > F
Course	1	60.84	60.84	5.70	0.0329
Experience	1	0.38	0.38	0.04	0.8531

Table 9.8 Contribution of Course and Instructor Experience

It may be that what course an instructor taught had more to do with their likelihood of saying that they would use QuesGen again than did experience. More results regarding these two independent variables will be presented next, and all of these results will be discussed further in the next chapter.

9.7 Instructor Experience and Course Differences

Hypotheses 6, 7, 8, and 9 addressed predictions related to differing levels of instructor experience with writing questions. Hypotheses 10, 11, 12, and 13 dealt with the effects associated with the different courses that were being taught. Results pertaining to hypotheses 6, 9 and 10 have been presented already. In the next two sections, the remainder of the results related to these two independent variables will be presented.

9.7.1 Different Amounts of Experience

H7 predicted that questions written by more experienced instructors would be more likely to be rated as being higher in Bloom's Taxonomy. 0 displays the result of a chi-squared analysis of the relationship between these variables. Experienced instructors were

significantly more likely to write questions that the judges perceived to be at higher levels in Bloom's taxonomy than questions written by inexperienced instructors, supporting the hypothesis. It should also be noted that even though experienced instructors were more likely to write higher-order questions, about five-eighths (63.25%) of the questions overall were rated at the recall level. If one accepts the premise that a high quality question for a college audience is one that targets cognitive skills higher than recall, then these results support the underlying argument of this thesis, that college instructors are not as competent as they might be at writing questions.

Instructor Experience Writing MCQs	Judge's Evaluation		Total
	Recall Level	Higher than Recall Level	
No Experience	67.5 84.38%	12.5 15.632%	80 40.0%
Experienced	59 49.17%	61 50.83%	120 60.0%
Total	126.5 63.25%	73.5 36.75%	200 100%
$\chi^2 = 51.197$ $p < 0.0001$, $N = 200$			

Table 9.9 Relationship Between Instructor Experience and Judges' Evaluations that Questions Assessed Higher-Order Bloom's Taxonomy Levels

H8 predicted that questions written by instructors with more experience writing MCQs are likely to have fewer technical flaws than questions written by more inexperienced instructors. To evaluate this question, a score was generated using items 3 through 23 on the item-review checklist (Appendix C), following the same method as described earlier in Section 9.4.3. Since as reported earlier these scores were not normally distributed, a Mann-Whitney U test was run to determine if a relationship existed between the variables. The data collected support this hypothesis ($U=27,264$, 1-sided $p<0.0001$, $N=200$).

9.7.2 Different Course Subjects

H11 predicted that there would be no difference in the likelihood that a question would be categorized as a higher-order Bloom question for different course subjects⁵. The chi-squared analysis shown in 0 indicates that this was not the case. Instructors in the GCOM course were significantly more likely to write questions that the judges evaluated as higher-order. Again, this analysis shows that while slightly over half of the questions developed by the GCOM instructors were rated to be higher-order, still the majority of questions (63.25%) are written at the recall level.

Course Subject	Judge's Evaluation		Total
	Recall Level	Higher than Recall Level	
GCOM	43.5 43.50%	56.5 56.50%	100 50.0%
GKIN	83 83.00%	17 17.00%	100 50.0%
Total	126.5 63.25%	73.5 36.75%	200 100%
$\chi^2 = 67.1238$ $p < 0.0001$, $N = 200$			

Table 9.10 Relationship Between Course Subject and Judges' Evaluations that Questions Assessed Higher-Order Bloom's Taxonomy Levels

H12 predicted that there would be no difference in the number of technical flaws for questions written for different course subjects. On the contrary, a Mann-Whitney U test indicated that GCOM instructors were significantly less likely to have technical flaws in their questions ($U=44,937$, 1-sided $p<0.0001$, $N=200$). This hypothesis was therefore unsupported.

⁵ Hypotheses of "no relationship" are somewhat unorthodox as research hypotheses. This hypothesis was made at the time the experiment was designed and therefore before it was known what courses would be involved in the study. Knowing ahead of time what courses would be involved would most likely have led to the hypothesis of a definite and directed relationship between the variables. Arguments for the existence of such relationships are made in the next chapter for all of the hypotheses 10 through 13 which predicted "no difference" between the groups being observed.

H13 predicted that there would be no difference in user satisfaction with QuesGen for different course subjects. A validated user satisfaction instrument (QUIS) was used to gauge users' satisfaction with the various components of QuesGen. Figure 9.3 below plots the results for all participants combined. The possible range for scores was from 1 to 9, with higher scores indicating greater satisfaction. The first observation of this data is that the average value, indeed the lower quartile value, is above five for all categories, indicating that satisfaction was higher than neutral.

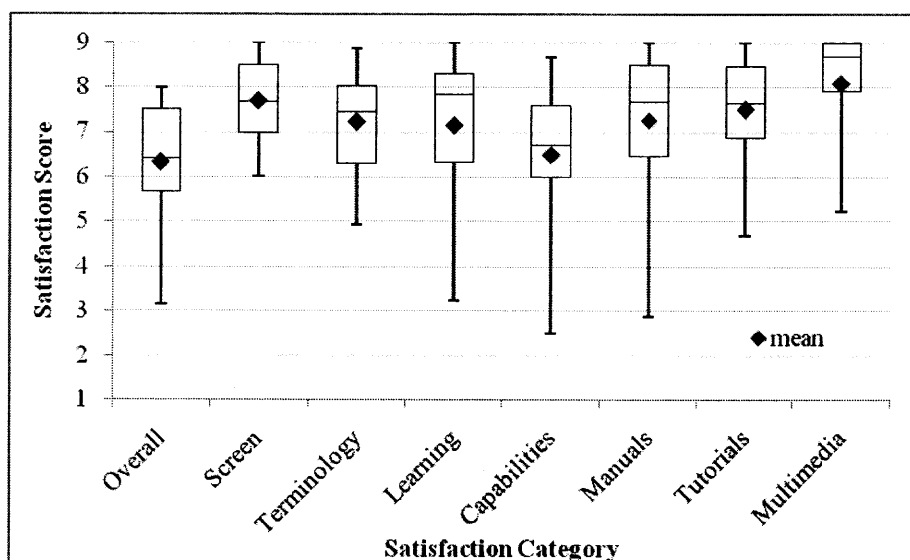


Figure 9.3 Satisfaction Scores for All Participants (N=17)

Shapiro-Wilk tests for normality indicated that data were normally distributed for all variables except for multimedia. The GKIN and GCOM satisfaction scores were compared using a *t*-test for all normally distributed variables and Mann-Whitney U test for multimedia. No significant differences were found between the course groups. Subsequently a power analysis was performed for the parametric tests to determine the likelihood of detecting differences. The results of the power analysis indicate that the sample size for this study was insufficient to detect any differences between these groups.

These results are summarized in 0. These data failed to provide evidence for rejecting H13, but given the low power of the tests, it is questionable whether or not a difference would have been found even if it existed.

Satisfaction Category	Test	<i>p</i>	Power ($\alpha=0.05$)
Overall	$t= 0.77$	0.4552	0.111
Screen	$t= 0.22$	0.8278	0.055
Terminology	$t= 1.14$	0.2729	0.196
Learning	$t= 1.13$	0.2777	0.184
Capabilities	$t= 0.53$	0.6062	0.078
Manuals	$t= 1.37$	0.1947	0.276
Tutorials	$t= 1.07$	0.3023	0.178
Multimedia	$U = 70.0$	0.8728	n/a

Table 9.11 Results of Significance Tests for Satisfaction Differences Between GCOM and GKIN Instructors

9.8 Students' Evaluations of the Questions

Hypotheses 14a through 16c all deal with students' perceptions of the difficulty, clarity, and fairness of the questions that the participating instructors wrote, each with respect to one of the independent variables—use of QuesGen, experience, and course. All of the students' evaluations were based upon 5-point Likert-type items. The scales were scored from 1 to 5 with 1 being “strongly disagree” and 5 being “strongly agree.” Therefore for the difficulty scale, a score closer to 5 indicated that students thought it was a more difficult question, whereas with the clarity and fairness questions, a score closer to 5 indicates that a question is clearer and fairer. A score of 3 indicated neutrality. An average score was calculated for each of the 200 questions written by the teachers based on the ratings of the students who took the quiz containing that question. As such the N for each of the measures reported below is 200, the number of questions. These averages were calculated based on 636 student responses to the follow up survey. This means that

there was an average of over 30 students responding for each of the 200 questions. In addition, students were asked to list any words that they didn't understand or know the meaning of, and also to indicate why they believed that they missed a given question, in the event that they did miss the question. These latter measures were used to give deeper insight into the analysis of individual questions, to help make arguments about the quality of those questions. Analysis of individual questions occurs in the next chapter.

9.8.1 QuesGen's Impact on Perceptions of Difficulty, Clarity and Fairness

Hypotheses 14a through 14c all dealt with the impact of QuesGen on students' perceptions and predicted that questions written by instructors using QuesGen would be perceived as more difficult, but also clearer and fairer than questions written without QuesGen. A Shapiro-Wilk test for normality found that the distribution of data differed significantly from normal in all cases. As such, Mann-Whitney U statistics were calculated, the results of which are summarized in 0. Since hypotheses indicated a direction, the one-sided approximation of Z was used to determine significance.

Student Perception	QuesGen Mean	Non-QuesGen Mean	U	<i>p</i>
The question was difficult	2.17	2.31	21,114	0.0027
The question was clear	4.31	4.26	26,038	0.0310
The question was fair	4.17	4.13	26,038	0.0987

Table 9.12 Results of Tests of QuesGen's Impact on Student Perceptions

H14a predicted that students would perceive questions developed with QuesGen to be more difficult than questions developed without. The reasoning was that these questions would be more likely to assess higher-order thinking, which students in turn would perceive as harder. Contrary to this expectation, a significant relationship was found in the opposite direction. This hypothesis was not supported. Students indicated

that the questions generated without QuesGen's new functionality were more difficult, although the mean scores for difficulty for both the QuesGen and non-QuesGen questions were well below three which means that on average students disagreed the questions were difficult.

H14b predicted that students would perceive questions written with QuesGen to be clearer than questions written without. The rationale for this was that following the question quality checklist would lead instructors to write concise, grammatically correct questions that contained audience-appropriate vocabulary. The results of this test indicated that to a small but significant extent, questions written with QuesGen were indeed perceived to be clearer than those written without. This hypothesis was therefore supported by the data.

H14c predicted that students would perceive questions written with QuesGen to be fairer than questions written without. The reasoning behind this was that if instructors used QuesGen to align their questions with specific educational objectives, students would be more likely to be prepared for the questions that they would see on the quiz, and hence rate the questions as fairer. To a very marginally significant extent ($p < 0.10$), students did, in fact, indicate that questions created with QuesGen were fairer, though only slightly so. Given the number of responses upon which these scores were calculated, this was construed as weak support for this hypothesis.

In summary, the use of QuesGen appears to have had a significant, positive effect on students' perceptions of MCQs written by instructors. These results will be discussed further in the next chapter.

9.8.2 Impact of Instructor Experience on Difficulty, Clarity, and Fairness

Hypotheses 15a through 15c all dealt with the impact of instructor experience on students' perceptions and hypothesized that questions written by more experienced instructors would be seen as clearer and fairer, but also more difficult. As with the data in the previous section, this data also was found to depart significantly from the normal distribution, and therefore nonparametric tests were used to test relationships. The results of the tests are summarized below in 0.

Student Perception	Experienced Mean	No Experience Mean	U	P
The question was difficult	2.14	2.46	39,996	<0.0001
The question was clear	4.34	4.17	26,158	<0.0001
The question was fair	4.24	3.98	24,090	<0.0001

Table 9.13 Results of Tests of Impact of Instructor Experience on Student Perceptions

H15a predicted that students would perceive questions written by more experienced instructors to be more difficult. The reasoning for this was that it was thought that more experienced instructors would be more likely to write questions that tapped higher-order thinking. In turn, it was thought that students would find higher-order questions to be more difficult. As it turns out, the reverse trend was found. Students perceived questions written by inexperienced instructors to be significantly more difficult than those written by experienced instructors. This hypothesis was unsupported.

H15b predicted that students would perceive questions written by more experienced instructors to be clearer. The reasoning for this was that experienced instructors would know better how to phrase questions in a way that was pitched at students' level of understanding. The data supported this hypothesis, indicating that

questions written by experienced instructors were perceived to be significantly clearer than questions written by inexperienced instructors.

H15c predicted that students would perceive questions written by more experienced instructors to be fairer. Again the reasoning here was that instructor experience would lead to the writing of questions more systematically aligned with instruction, which in turn would be perceived by students as fair. Again, the data supported this hypothesis. Questions written by experienced instructors were perceived to be significantly clearer than questions written by inexperienced instructors.

Instructor experience appears to have had a strong impact on students' perceptions of questions. As mentioned before, this impact may be confounded with the impact of the course material, the effects of which are reported next.

9.8.3 The Impact of Course Content on Difficulty, Clarity, and Fairness

Hypotheses 16a through 16c all dealt with the impact of course content on students' perceptions and hypothesized that questions written for different courses would have different levels of perceived difficulty, clarity, and fairness, but not knowing prior to running the study which courses would be involved, hypotheses were not specific to the actual courses that participated. This data was found to depart significantly from the normal distribution, and therefore nonparametric tests were used to test relationships. The results of the tests are summarized below in 0. The only difference between this table and the ones in the previous sections is that a 2-sided test was used since the relationships were not hypothesized to exist in a specific direction.

H16a predicted that students would perceive questions written for different courses to have different levels of difficulty. The data supported this hypothesis.

Students perceived the questions in the kinesiology course to be more difficult than the questions in the communications course.

Student Perception	GCOM Mean	GKIN Mean	U	P
The question was difficult	2.13	2.40	24,090	<0.0001
The question was clear	4.28	4.26	41,170	0.3549
The question was fair	4.21	4.06	44,958	<0.0001

Table 9.14 Results of Tests of Impact of Course Content on Student Perceptions

H16b predicted that students would perceive questions written for different courses to have different levels of clarity. This hypothesis was not supported. There were no significant differences in the perceptions of clarity of the questions in the two courses.

H16c predicted that students would perceive questions written for different courses to have different levels of fairness. This hypothesis was supported. The students found the questions written for the communications course to be significantly fairer than the questions written for the kinesiology course.

9.9 Summary

This chapter presented the data and results of the statistical tests on those data with respect to all of the hypotheses of the QuesGen study. The next chapter will look at these results in more depth and make an effort to interpret them and decide what, if any, conclusions can be drawn.

CHAPTER 10

DISCUSSION

10.1 Improving Question Quality with QuesGen?

QuesGen was designed to help teachers write better multiple-choice questions. Chapter 6 went into depth operationalizing the concept of question quality with the intent that these constructs could then be used to gather evidence that would shed light on the question of QuesGen's effectiveness. So, was QuesGen effective? Were the questions developed using QuesGen's new features really "better" than those developed without? Were the differences in question quality due to QuesGen overshadowed by other more significant variables like instructor experience or course content? Were the constructs of question quality developed to measure the system meaningful and useful? Was the experiment that was conducted successful in answering the questions that it posed? This chapter is a discussion of these questions, and it will be organized around analyses of the hypotheses that were unsupported or turned out other than expected.

The first subject to be dealt with will be why QuesGen didn't seem to foster greater alignment of questions with the educational objectives targeted by instruction. Following that is an analysis of the item-review instrument used by the expert judges. The instrument didn't perform as well as was hoped in providing insight into question quality, and this chapter presents the results of an exploratory factor analysis that was performed to try to glean some more insights from the instrument. This factor analysis will also assist in the next section of this chapter, which looks at the problematic discrimination index (DI), and tries to make sense of why all of the hypotheses related to

DI were unsupported. The fourth section of this chapter will attempt to disentangle the large effects of instructor experience and course content which were associated with the most significant relationships to be observed in the study. The final section of this chapter will attempt to answer the question of QuesGen's overall effectiveness.

10.2 Why QuesGen Didn't Foster Alignment of Questions to Objectives

In the early stages of this research, the researcher was examining questions that had been written for several quizzes and had the opportunity to ask the instructor, "What was the point of this question?" The instructor was asked to work backward from the question and explain the underlying, salient educational objective being addressed by the question. Too often, it seemed, the answer was some variation of "Hmmm, I'm not sure," or "Well, I guess that's not too important." Experience as a test-taker, and with the types of questions included by textbook publishers with their texts led to the belief that encouraging instructors to align their questions with well-thought out, salient objectives would greatly improve the questions. This was the motivation for incorporating the feature in QuesGen that allowed instructors to select an objective from a dropdown list. The results of the study, however, showed that selecting an objective from a list does not seem to have an impact on the degree to which a question is perceived to be in alignment with stated objectives. Furthermore, even without QuesGen's feature, instructors were fairly likely to have written a question that addressed a stated objective. Why was this the case? Several possible explanations for this result will be discussed.

One explanation is that there was not enough content being evaluated. Control over the content of instruction was part of the study design. The purpose was to remove

individual differences in instruction as a variable in how students received the material and responded to the quizzes. Also it was desired to have instructors within a course all write questions about the exact same content. The GCOM lecture was 21 minutes long and addressed four objectives. The GKIN lecture was 12 minutes long and addressed five objectives. It is arguable that a ten question quiz about such a small unit of instruction is unrealistic unless the content is central to future instruction and/or difficult for students to absorb. Neither argument can be made for the content in these lectures. By restricting the content to such a great degree, and then explicitly asking the instructors to write ten questions about it, no latitude was left for the instructors to write questions about anything other than the objectives. There was not a lot, if any, extraneous or unimportant information in the video lectures that were used. As such, there may not have been enough content for instructors to focus on topics that were not salient.

A second problem is that in more realistic settings, explicit objectives are less likely to exist as they did in the QuesGen study. In the follow-up interviews, instructors were asked how often they defined explicit objectives for each unit of instruction, and how often they referred to them when writing questions. None of the instructors interviewed indicated that they wrote objectives on a unit by unit basis. If objectives existed, they were defined at the course level. In the QuesGen study, instructors were asked to write 10 questions about 4-5 objectives which were given to them on PowerPoint slides. Coupled with the knowledge that their questions were going to be analyzed, it is likely they worked harder to write good questions. In this situation, they simply looked at the slides and built questions directly from them. This would be a case

of Hawthorne effect, in which the instructors all work harder because they know they are being watched.

A third reason that a high proportion of the questions addressed objectives, and that there was little difference between the QuesGen and standard groups is that the objectives themselves targeted fairly low level skills. Since the lectures used in the courses came from standard syllabi, the experimenter didn't feel it was appropriate to step in and ask the instructors to change their objectives. The experimenter was not a subject-matter expert in either kinesiology or communications, and didn't feel it was his place to change the curriculum for core courses in the fields of the instructors who were participating. It was noted by the investigator that the objectives for the two courses contained overlap, and also focused overly much on factual information that could be learned by memory alone and didn't require any higher order processing on the part of the learner. Since left to their own devices, instructors tend not to write questions that address higher-than-recall cognitive skills, if the objectives only address recall-level issues, then it is likely that the instructors' questions will line up with them.

The quality of objectives that were created for the video lectures in this study seems to have been a weakness in the study design. At one point in the design of the research study, it was proposed that instructors would be able to write their own objectives, and enter them into QuesGen. This approach, it was felt, would introduce too much variability—i.e. the quality of the objectives, and the instructor's ability to write those objectives would be highly correlated to the quality of the questions that were written. However, it is clear in hindsight that more effort should have been spent on working with the instructors who created the video lectures to create objectives that

targeted higher-order cognitive skills. In addition, it may be a good idea to expand the amount of content that is addressed for the quizzes. It is plausible that if this had been the case, instructors using the objective-alignment feature of QuesGen would have been more likely to have the judges rate their questions as aligned with stated objectives than instructors who did not have or take advantage of this feature.

What is the impact of the use of QuesGen's dropdown menu for objectives on question quality? The results of the study indicate that this function contributed little to the relative quality of questions. However, given the observations discussed above, it is not possible to say that the feature is completely useless and should be left out of future implementations of the system. In particular, discussion to follow below with respect to the quality of objectives will bring stronger rationale for keeping it.

10.3 What the Item-Review Instrument Says About Question Quality

The item-review instrument (Appendix C) used by the expert judges to evaluate the questions was developed from the empirical literature on technical flaws with MCQs. Even though there was a high degree of agreement between the judges on the ratings given to the questions, this was mostly because there doesn't appear to have been a lot for them to disagree about. While there is a large number of potential question flaws, the number of flaws that appear in any given question appears to be few. Overall, therefore, the resolution of the item-review instrument was poorer than hoped for. This section discusses an exploratory factor analysis that was performed on the item-review instrument data.

After the initial results of the item-review instrument were examined as a whole, the response patterns on the individual items in the item-review instrument were examined more closely. The correlation matrix for all of the items on the instrument was computed (see 0 at the end of this chapter). It was found that five of the best practices were violated by almost no one. As such variance on these items was close to zero and they were removed from the scale since their presence shed little light on question quality. The five items that were removed were:

- The vocabulary in this question is appropriate to the student level
- The question inappropriately uses humor (reverse coded)
- This question follows best practices in the use of the words *NOT* and *EXCEPT*
- This question avoids using “all” or “none of the above”
- This question avoids using absolutes such as “every” or “never”

One of the first things to appear from the correlations was that the items worded “The item is concise” on the item-review instrument was not significantly correlated with any of the other items. In 400 ratings the judges rated items as not being concise only fourteen times (average score 1.9375/2, variance=0.11). One possible reason for this was that the questions really were concise. However, inspection of the questions seems to indicate that there is a fair amount of variation in the length of the questions. Another possible reason is that from the judges’ perspective, most of the questions were appropriately concise—meaning that given what the questions were trying to accomplish, they were not overly wordy. The students’ evaluations seemed to agree with the judges’ analysis, at least in the case of wordiness—the students ratings of “clarity” were significantly lower ($\chi^2=6.73$, $DF=2$, $p=0.0346$) when the judges indicated that questions

were not concise. The main independent variables in the study (use of QuesGen, instructor experience, and course) were not related to conciseness (chi-squared tests did not indicate any significant relationships); however, out of the forty questions where instructors actually used the question-quality checklist, none of the questions were rated as not being concise. This was significantly different ($\chi^2=11.93$, $DF=2$, $p=0.0026$) from the instructors who did not use the checklist, but given the small number of questions rated as not concise (only fourteen out of 400), this result may not be reliable. This variable was removed from further analyses.

Item	Factor1	Factor2	Factor3
1a	0.81771	0.07362	0.10129
1b	0.67327	0.13845	0.07333
1c	0.56742	0.02273	0.20885
1d	0.51789	-0.01049	0.27333
1e	0.10834	0.09285	-0.01404
2a	0.08065	0.74837	0.17124
2b	-0.04034	0.72028	0.05599
2c	-0.02065	0.58842	0.03972
2d	0.1613	0.44464	0.01932
2e	0.06735	0.11545	0.09157
3a	0.11165	0.10281	0.69697
3b	0.06722	0.12454	0.48629
3c	0.2082	-0.02557	0.32153
3d	-0.01445	0.00083	-0.44179
3e	-0.1505	-0.11349	-0.59072

Table 10.1 Factor Loadings for Varimax Rotation of Item-Review Instrument Items

After the above variables were removed, an exploratory factor analysis was run on the remaining fifteen variables and yielded three factors. The factor loadings calculated using a varimax rotation are shown in Table 10.1. The factors are phrased as questions with the items that were included in the factors listed below them in order by contribution to the explanation of variance.

1. Is the answer correct?

- a. The question has exactly one correct answer.
 - b. The answer marked as the key is the best answer.
 - c. This is a trick question (reverse coded).
 - d. The content in this question is based upon opinion (reverse coded).
 - e. The question is grammatically correct.
2. Are the distractors of high quality?
- a. All of the answer choices for this question are plausible.
 - b. The answer to this question is not obvious.
 - c. The question's distractors are based on likely student misconceptions.
 - d. The answer choices in this question use parallel sentence structure.
 - e. The content assessed in this question is trivial (reverse coded).
3. Is the content well-conceived?
- a. This question addresses exactly one educational objective.
 - b. This question clearly addresses a stated educational objective.
 - c. This question depends upon cultural knowledge (reverse coded).
 - d. The concept examples used in this question are novel.
 - e. Students can answer this question purely from memory (reverse coded).

Mann-Whitney U tests were run to assess the relationship between the independent variables in the study and these factors. Additionally, the QuesGen interface group was further restricted to just the subset of questions for which the instructor had used the question quality checklist, since it was thought that use of the checklist might be related to high performance on the item-review instrument. The results of these tests are presented in Table 10.2. None of these variables were associated with factor two ($p > 0.05$)

for all tests). Both instructor experience and course (which are highly related to one another) were strongly associated with factor three.

Variables	Factors		
	Factor 1	Factor 2	Factor 3
Used QuesGen	U=25,457, $p=0.038$	U=24,273, $p=0.407$	U=24,394, $p=0.369$
Used Checklist	U=8760, $p=0.076$	U=8454, $p=0.233$	U=7923, $p=0.442$
Experience level	U=30,975, $p=0.095$	U=32,207, $p=0.448$	U=24,751, $p<0.0001$
Course	U=40,648, $p=0.262$	U=39,661, $p=0.329$	U=48,220, $p<0.0001$
U = Mann-Whitney U, 1-sided, N=200			

Table 10.2 Results of Tests for Relationships Between Factors and Independent Variables

Are these relationships meaningful? What, if anything, do these independent variables have to do with these factors? In the case of Factor 1, the use of QuesGen was associated with higher scores on the factor. This would indicate that QuesGen users were less likely to write ambiguous questions (exactly one correct answer), less likely to miskey their questions, and write questions that are tricks, based upon opinion, or grammatically incorrect. Since QuesGen was designed to help teachers acquire these exact skills, it is reasonable that use of QuesGen should be associated with these things. On the other hand, QuesGen was designed to promote many other best practices with which its use does not appear to be related. It is not clear why use of QuesGen would be associated with this particular set of question quality improvements and not others.

Factor 3 was significantly related to both instructor experience and to course content. In this case, more experience and teaching GCOM were associated with higher levels of the factor. This would indicate that more experienced teachers, and teachers who teach GCOM rather than GKIN, are more likely to write novel questions that address exactly one stated objective, can't be answered purely from memory, and don't

rely on cultural knowledge to be answered correctly. Since there is an overlap between experience and course, it is likely that both of these factors contribute to the relationship, and there are reasons that would explain both. First, it is reasonable to expect that a more experienced teacher is going to have a greater wealth of experience and examples to draw upon when writing a question, and as such will be more likely to write questions that are more novel. As for the course influence, the stated objectives for the GCOM lecture say that students should be able to identify power resources within a “communications scenario.” This wording suggests that instructors should come up with scenarios, which is indeed what they did. It is not possible to figure out whether experience or course contributed more to the relationship, but both variables would tend to increase the strength of the relationship and explain its significance. As before, however, we are left wondering why instructor experience did not lead to a significant relationship with Factors 1 and 2. Since there is not as compelling an explanation for why course content would be associated with Factors 1 and 2, it is possible that it was the combination of the two variables in the case of Factor 3 which made the difference.

More will be said about this factor analysis later in the discussion of the discrimination index, which comes next.

10.4 Interpreting the Discrimination Index

In Chapter 6 it was argued that DI is a good quantitative measure of question quality for three reasons. First, it is easily calculated, which can't be said for any other measures of question quality. Second, given that group-dependency and test-dependency are controlled for, it is internally valid. Third, the DI has face validity since the goal of the

DI is to divide students into groups who have and have not mastered the content addressed in the question. Knowing how many and who these students are would aid the instructor in designing instruction and/or remediation for students who have not yet mastered the content.

Of the three hypotheses made with respect to DI, two were unsupported and the third showed a significant result in the opposite direction of that predicted. This result suggested to the investigator that his original conception of DI needed serious revision, and prompted a deeper look at DI as a measure of question quality. Correlations with the results of the expert judges' evaluations with the item-review instrument were used to develop an argument as to why the results came out as they did. Three observations came out of this analysis. First, DI has almost no relationship to an instructor's likelihood to write a question tapping higher-order thinking. Rather, the DI seems to have more to do with the quality of the distractors that were written, which is the second observation. Third, the DI may be more sensitive to the degree to which a question assesses pure recall, which limits its usefulness for assessing higher quality MCQs which, as defined earlier, address cognitive skills more complex than memory. These results and analyses will be presented now.

The analysis that was performed was to correlate the DI scores with the factors that were generated by the factor analysis presented earlier. Table 10.3 shows the correlation coefficients and their significance levels between DI and the three factors. The first interesting thing to note is that DI is significantly correlated with Factor 2, which was earlier labeled as asking the question "Are the distractors of high quality?" In

the earlier analysis, it was shown that none of the independent variables was significantly related to Factor 2. An intuitive argument as to why DI and Factor 2 are related follows.

	Factor 1	Factor 2	Factor 3
DI	0.02371	0.12882	-0.12225
	0.6364	0.0099	0.0144

Table 10.3 Correlation Matrix of DI with Factors from the Item-Review Analysis

What are the ways that distractors can be poorly formed, and what are the types of problems that arise because of this? Poorly written distractors are implausible, contain grammatical errors, give themselves away with incongruent grammatical constructions, and in general fail to distract test takers away from the correct answer. The fewer effective distractors an item has, the lower the DI, since the chance of guessing correctly goes up and there is less likely to be a difference between the “high ability” and “low ability” groups. Therefore it makes sense that if Factor 2 measures the quality of distractors that higher Factor 2 scores would indicate greater numbers of effective distractors, which would in turn lead to higher DI scores. Likewise, lower scores on Factor 2 would correspond to lower DI scores.

To test this theory, the distractors of all of the questions that had a DI of zero were examined. Recall that a DI of zero results when all of the upper and lower quartile test takers answered correctly in exactly the same proportion. Usually this means that either everyone missed the question, or everyone got it right. There were 22 questions written for this study that had a DI of zero. Of those, 20 out of 22 had a zero DI because everyone answered correctly. Out of the 20, 9 were written by GCOM instructors, 11 by GKIN instructors. Furthermore, 6 of these questions were written by inexperienced MCQ writers, and 14 were written by more experienced instructors. What is important to

call out in this analysis is that almost all of the distractors for these questions were completely ineffective at getting test takers to select them. A first assumption might be to say that the questions were all too easy, as this one which was answered correctly by 100% of the 22 students who answered it on their quiz:

- Q: Which of the choices below is not a macronutrient?
- A. Carbohydrate
 - B. Protein
 - C. *Minerals***
 - D. Fat

This question violates the best practice of having a non-obvious correct answer among the distractors. This question targets the absolute most basic bit of information in the video lecture. It is so unlikely that students will miss this question, that the information the instructor stands to gain by asking it is predictably going to be very minimal.

However, not all of the questions that had zero DI were of this form. Some were quite complex, as was this one which was answered correctly by all 24 of the students who encountered it:

- Q: Gwen, a student in organizational communication, contacted a local company in Harrisonburg, Comsonics, and asked to sit in on its executive board meeting to study relational messages between employer and employees. Because she is 15 minutes late to the appointment, she is not allowed into the room but can observe participants through a one-way mirrored window. Having never met the company's board members, she is trying to figure out who the president of the company is. What nonverbal indicators should she follow?
- A. *Clothing, eye contact and touch should indicate who has the most power in the group.***
 - B. She should try to see who speaks to whom the most
 - C. She should apologize and make an appointment to come back some other time
 - D. Because she cannot hear members of the executive board's discussions, she cannot assess who has the most power in the group

This second question was rated to be at the analysis/comprehension level by the expert judges, yet when one analyzes the distractors closely, it becomes apparent that out of the four only one of them actually even mentions any nonverbal indicators of power. It is hard to say that a test-taker needs to be able to comprehend or apply the concepts related to nonverbal communication in order to answer this question correctly. The answer choices are not parallel. Two are actions: “She should....” The fourth choice is a muddled way of saying “none of the above.” Only the correct choice is even of the format one would expect for a correct answer. Better distractors would have followed the same pattern as the correct answer and mentioned potential forms of nonverbal communication that could be watched to determine the power relationships in the group. This is a situation where poorly constructed distractors led an otherwise creative and potentially informative question to yield little information about what the students actually know about nonverbal communication.

On the other end of the spectrum, here are the two questions that had the highest DI scores, 1.0 and 0.91 respectively:

Q: What is the appropriate recommendation for fat?

A. 10-20%

B. 15-30%

C. **20-35%**

Q: Fats provide _____ calories per gram.

A. 3

B. 4

C. **9**

D. 10

Both of these questions rely purely upon students' ability to remember the right answer from having seen it in the lecture. The first question was answered correctly by 24/39 (63%) of the students, with 11/39 (29%) choosing A, and 4/39 (11%) choosing B. Over

25% of the students who responded to the survey thought the first question was unfair, and 18% thought the wording was not clear. The second question was answered correctly by 23/41 (58%) of the students with 8/41 (20%) answering A or B, and 2/41 (5%) answering D. The distractors in these questions follow best practices in that they are parallel, non-obvious, plausible, based on likely student misconceptions, and ultimately effective at getting students to select them. There are no contextual, grammatical, or language clues that would give away the correct answer. Numeric answers are sorted in ascending order. It is relatively clear why these questions have high DI scores—the only way to answer these questions correctly is to have remembered the content from the video lecture. It is likely that the students who scored most poorly on the quiz either didn't watch, didn't pay close attention, or simply didn't retain the information in the video lecture. It should be relatively unsurprising that the “better” students in a class are more likely to pay attention to lectures and retain the content therein. These questions have value in that they clearly call out the 37% and 42% of students who watched and remembered the content from the lectures.

The obvious flaw with both of these questions, on the other hand, is that they don't ask students to demonstrate understanding, or show that they can apply their knowledge of fat to the health-related topics discussed in the lecture. These questions do not communicate the value or utility of understanding these basic facts about fat. They reinforce that “learning” is more or less equivalent to being able to memorize and regurgitate facts that have little meaning in the context of the students' lives. The instructor has missed an opportunity to connect information about fat to situations in which students might apply that knowledge.

What do these example questions illustrate about the value of the DI for evaluating question quality? First, the analysis of these questions and their distractors lends credence to the association of the DI and Factor 2 from the factor analysis, whose questions are related to the quality of question distractors. It appears from this analysis that a very low DI score is a good indication that there are problems with all or most of the distractors. Second, they serve to highlight that the DI is *not* good at identifying the degree to which questions address higher-order cognitive abilities since the questions seemed equally likely to be lower or higher on Bloom's taxonomy regardless of having a high or low DI. These two conclusions will be reinforced by the analysis of the other significant correlation shown in Table 10.3.

Table 10.3 indicates that the DI was negatively correlated with Factor 3 from the factor analysis shown earlier. This factor was labeled "Is the content well-conceived?" and the questions in the factor addressed how closely the question was aligned with objectives and also how likely the question was to ask purely recall or memory questions. Questions that address only memory have lower scores than questions that address higher-order thinking. Questions aligned with objectives score more highly. The analysis of questions with extremely high DI scores suggests that such questions are very likely to ask students to be able to remember minute details from the content of a lecture—details that they have no way of guessing through context clues. The association of high DI scores with purely memory questions is consistent with a negative correlation to factor 3, and reinforces the notion that while DI may indicate the degree to which distractors are well constructed, a high DI is not necessarily "high quality" when viewed from a cognitive perspective. To be sure, a high DI is desirable insofar as it allows an instructor

to separate students into groups based on concept mastery, a DI that is “too high” may indicate that the question has a tendency to focus on the assessment of rote memory.

In summary, the analysis of the DI scores makes some important contributions to this study. First, the DI score analysis provides actionable insight into the meaning and interpretation of extremely high and low DI scores. In extreme cases, the DI scores suggest that the distractors of a question are very well or very poorly constructed. This insight highlights good distractor design as one of the core elements of good question design. That the use of QuesGen was correlated neither with DI scores nor with Factor 2 from the item-review instrument is an indication that the next version of QuesGen needs to take a different approach to addressing the issue of distractor quality.

10.5 Course and Experience Effects

Some of the most significant differences in question quality were seen between the different groups of instructors. As already indicated in Section 9.3, which discussed limitations with the study, there is a large overlap between the breakdowns of instructors by course and experience. All of the inexperienced instructors happened to fall into the GKIN group, and almost all of the experienced instructors were in the GCOM group.

Question Writing Experience	Course	
	GCOM	GKIN
None	0	8
At least 1 semester	10	2

Table 10.4 Distribution of Instructors by Course and Question Writing Experience

As can be seen from the table, all of the teachers in GCOM had some experience in writing MCQs, and most had been writing them for several years. Only one person indicated that this was his second semester writing questions. On the other hand, eight

out of the ten instructors in GKIN were graduate assistants who had never written an MCQ before participating in this study. Since there was such a strong correspondence between experience and course, it may be difficult to do other than offer conjecture about which variable accounted for the greater share of the variance. What are the arguments that could be used to explain the impact and interaction of these two variables?

First, what are the arguments that would suggest that course content has a larger effect than experience? One argument would be that if the course content taught by one of the groups was somehow related to the field of educational assessment, that group might have an inherently greater likelihood to write questions that would be seen as high quality. Since one way to view assessment items in general, and MCQs in particular, is as a communication tool between instructors and students, it is arguable that the professors teaching GCOM (Fundamentals of Human Communication) had an advantage over the GKIN professors in that the core concepts of their field have something in common with the field of educational assessment. It is difficult to say, however, that this potential overlap between the fields will be a more powerful determinant of question quality than the raw experience of an instructor.

Another reason that course content may be the stronger determinant of question quality is related to the objectives that were listed and their relationship to the ways in which question quality was measured. As discussed earlier, the educational objectives (see Figure 10.1) that were listed in the slides for the two courses were not of very high quality. Their primary weakness was their failure to balance definitions and other memory goals with higher order application or analysis goals. For example, an objective that would have fit with the lecture on nutrition might have been: Apply knowledge of

acceptable macronutrient ranges to scenarios in which people must make dietary choices. Since the experimenter was not a subject matter expert in either communications or nutrition, and since both of these courses have common syllabi that a great number of instructors work from, it was not deemed appropriate to suggest to the participants in charge of creating the video lectures that they revise their objectives. Some of these objectives suggest to instructors the creation of novel scenarios, such as GCOM #4. Based on these objectives, it is reasonable to hypothesize that GCOM instructors would be more likely to include novel scenarios in their questions (which they did), and in turn that judges would be more likely to rate their questions highly in Bloom's taxonomy because of this novelty (which they did).

<p>GCOM Objectives</p> <ol style="list-style-type: none">1. Define power and give an original example in your own life.2. Define and identify the 3 forms of power.3. Define and identify the 5 types of power resources.4. Use the terms to label which power forms and resources are present when presented with a communications scenario. <p>GKIN Objectives</p> <ol style="list-style-type: none">1. List the six essential nutrients and describe their functions in the body.2. List the acceptable macronutrient distribution ranges.3. Explain the difference between complete and incomplete proteins.4. Explain the difference between saturated, unsaturated, and trans fat.5. Explain the difference between simple and complex carbohydrates.
--

Figure 10.1 Educational Objectives from the Two Courses

This line of reasoning and the results of the study reinforce the value of high-quality objectives as important to the creation of high-quality MCQs. Although not

conclusive by any stretch, these results suggest that a future study would do well to try to isolate the effects of explicitly stated, high quality objectives as a determinant of the quality of the resulting questions. As originally conceived, users of QuesGen were to be responsible for generating their own objectives, and tutorials were developed to show them how to do this. This effort was dropped because of the added complexity it added to the study, but it should be picked up again in the future.

Second, what are the arguments that would suggest that experience writing MCQs would be a more powerful determinant of question quality than the field of expertise of the instructor? Intuitively, if the field of expertise of the instructor is not one that overlaps with educational assessment, then all other things being equal, one would expect that the more experienced instructor would write better MCQs. If, for example, a biology professor and a history professor, the former having four years of experience writing MCQs and the latter having only one year of experience, were to be compared, one would probably not argue that their fields would be the dominant factor in determining the quality of the questions that they would write.

It is natural to expect that MCQ writing, as with most human activities, is a skill that gets better with practice. However, given the quality of even experienced instructors, there seems to be a quality ceiling above which instructors do not rise without some form of training. A hypothetical graph showing this trend is depicted in Figure 10.2. “Training” in this figure is broadly conceived in this scenario to include anything from taking courses, to following the tutorials built into a tool like QuesGen. Training fosters reflection on practice and an effort towards an improvement in quality. In this study, the

GKIN instructors were at origin of this graph, whereas the GCOM instructors fell somewhere along the “without training” s-curve.

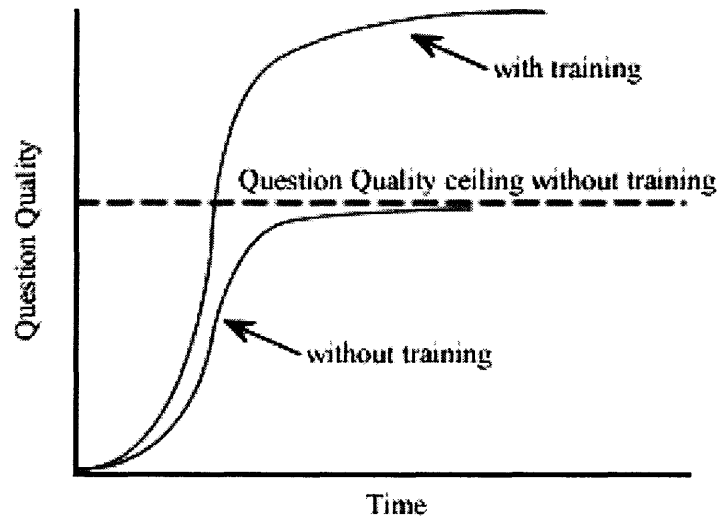


Figure 10.2 Hypothetical Graph of Improvement in Instructor MCQ Quality Over Time With and Without Training

Related to the notion that instructors get better with experience, an interesting phenomenon was observed during the follow-up interviews. Instructors do not typically analyze their students’ performance on MCQ-based tests. The MCQ portions of tests are frequently delivered using Scantron sheets or some other form that allows for automated scoring in which the instructor never has to see or think about the answer patterns for any particular question. In addition, since instructors want the opportunity to reuse questions in the future, they collect the test papers from the students, so the students never see the questions again after they have taken the tests. Teachers report that students will argue for points back on a test where they scored poorly, but seldom request to see the questions again in order to figure out what questions they missed. This combination of events means that instructors are likely never to know when they have written a bad question. This lack of reflection on the quality of one’s assessments would prevent an

instructor from learning through experience. After being shown the Question Analysis Reports (see a sample QAR in Appendix D), and having the QAR explained to them, all of the instructors expressed an interest in having such feedback provided to them in the future. When it was explained that this would mean delivering the follow-up survey to their students after every quiz or test, several said that perhaps they would only use such a tool occasionally, but several said they would use it every time.

In summary, even with these explanations for why either course content or instructor experience is likely to be the stronger determinant of question quality, it is still not possible to really separate out these effects with the data collected. This overlap between these variables was something that could not have been foreseen or controlled for in this experiment. Future studies should be able to deal with this more effectively.

10.6 Is QuesGen Effective?

The discussion of question quality in Section 6.2 ended with this question:

Does this measure provide insight into aspects of a question that have an immediate impact on the teacher-student relationship, and the student learning that results thereof?

This question was written to guide the interpretation of all of the measures of question quality to be gathered in this study. The primary measures used to evaluate the quality of questions were: Bloom's taxonomy level, expert judges' evaluations, the DI, and student evaluations. This section will further deepen the discussion of what it means for a question to be of high quality, and, in light of the data, make a determination of whether or not QuesGen was effective at achieving its goal of helping teachers write better multiple-choice questions.

Does knowledge of where a question falls in Bloom's taxonomy have an immediate impact on the teacher-student relationship, and the learning that results thereof? Low-level questions only assess whether or not a student remembers a given fact or figure at a given point in time. Such questions do not give any indication of whether or not the student understands the significance, value, or application of that bit of information. Low-level questions allow students to regurgitate knowledge unreflectively. One could train a parrot to respond appropriately when asked for the name of the first president of the United States, but that doesn't indicate that the parrot knows the significance of this information. If the parrot can identify the name of the first president, does it mean that the parrot is now ready to move on to learning the significance of George Washington's decision to step down after only two terms in office? If this question seems nonsensical, it is meant to illustrate that low-level questions may not provide enough information for teachers to make pedagogical decisions about how and what topics to cover next with students.

Higher-order questions, on the other hand, require not only that students recall facts and information, but also that they be able to apply that information in new and different contexts. Does students' knowledge of Washington's precedent-setting two terms allow them to see the significance of FDR's being elected four times? Does it allow them to understand the nature of power and why we now only allow our presidents to server at most two terms? This is the type of information that higher-order questions can elicit and it is the kind of knowledge that teachers need to make pedagogical decisions. The answer then is yes—knowledge of where a question falls in Bloom's taxonomy does have an impact on the teacher-student relationship. Therefore a measure

of QuesGen's success is the degree to which it helped teachers write questions that were higher order.

Did use of QuesGen lead to the writing of higher-order questions? Teachers who used the question templates built into QuesGen were significantly more likely to write questions that assessed higher-order thinking. *If* teachers use this feature, then it appears that QuesGen is effective at improving question quality in this dimension. A problem is that most of the teachers who had access to the question templates chose not to use them.

Do expert judges' ratings have an immediate impact on the teacher-student relationship, and the learning that results thereof? A person who has been trained to spot question flaws can spot questions that are likely to be unhelpful in identifying whether or not students have attained sufficient mastery in the subjects being studied. Catching flaws before questions have been delivered can give instructors time to adjust their questions for maximum effectiveness. Again, good questions inform sound pedagogical decisions on the part of the instructor. The answer to the question is yes—expert judges' ratings can have an immediate and positive impact on the teacher-student relationship. If QuesGen were to provide a mechanism that allowed the questions to be evaluated in the same way an expert judge would, it could improve question quality.

While it is unrealistic to have judges waiting to review every question that a teacher writes, it is not terribly difficult for a teacher to be trained to spot the flaws that an expert looks for. This was the motivation behind including the question quality checklist in QuesGen. The items in the item-review instrument were essentially the same items in the question quality checklist. Unfortunately, in the form that they were presented, instructors who used QuesGen did not use the checklist, and therefore the use

of QuesGen was not related to a reduction in question flaws. As such, QuesGen was not effective in this dimension of question quality.

Does knowledge of the discrimination index (DI) have an immediate impact on the teacher-student relationship, and the learning that results thereof? Analysis of the DI results from the study indicated that the DI can be useful in identifying problems with item distractors. Since the DI can only be calculated *after* students have taken a quiz or test, the DI itself cannot be used by a teacher to avoid delivering subpar questions to the students. However, since it can be easily and automatically calculated based on the results of a quiz, the DI can be used by instructors to avoid a misdiagnosis of misconceptions held by the students about the material being studied. Since the DI can be used prior to returning feedback to students on the results of a quiz, the DI can have an impact on pedagogical decisions and the teacher-student relationship. Therefore DI can be used as one indicator of question quality. Results from the study show that DI scores that are either too high or too low indicate problems with distractors. As such a measure of QuesGen's effectiveness would be that it led to the production of questions with mid-range DI scores.

There was no difference in the DI on questions between those developed with and without QuesGen. In that respect, QuesGen was not effective at increasing question quality. The item-review instrument, particularly Factor 2, indicated a set of questions that are related to distractor quality. Enough information was gained from the study to make some concrete recommendations about the redesign of QuesGen. The question quality checklist was either too intimidating because of its length, or too easily ignored, or perhaps both. A way needs to be found to make it less intimidating, and more

assertive about the way that it suggests changes to questions. Also since a large number of the checklist guidelines didn't seem to apply in many cases, some sort of pattern recognition that would enable the interface to "intelligently" notify the instructor of potential question flaws seems desirable.

Finally, does knowledge of students' evaluations of questions have an immediate impact on the teacher-student relationship, and the learning that results thereof? As with the DI, student's evaluations of questions can't occur until *after* the questions have been delivered, but student feedback is clearly relevant to the interpretation of question quality. In the QuesGen study, one instructor wrote a question that asked students to identify a "hedge" as a verbal indicator of power. About eight of the students indicated that they didn't know what the word "hedge" meant in the context of the question, and the instructor realized that this term had not been covered in the lecture. This is a serious flaw that is very unlikely to be found in any other way than through student feedback. However, given the trend identified in the follow-up interviews of neither students nor instructors reviewing the questions after a quiz, it is very likely that such errors are going unnoticed on a regular basis. Also in the study, there was a significant relationship between the questions judges rated as verbose and questions students marked as unclear. A third result from the student feedback was the occurrence of the selection of "careless error" as the reason students missed questions. Careless error indicates that students actually knew the correct answer, but for some reason clicked on the wrong button. This causes problems for interpretation of question results. All of these results indicate that yes, students' evaluations of questions are meaningful indicators of quality. As such an indicator of QuesGen's effectiveness would be if use of QuesGen was associated with

students rating questions to be clearer and fairer. This was indeed the case. Use of QuesGen was significantly related to students' rating questions as clear and fair.

In summary, QuesGen was effective in some ways, and ineffective in others. QuesGen's template feature is associated with the writing of higher order questions, which in turn have a greater potential to inform pedagogical decisions. Questions written with QuesGen were more likely to be rated as clear and fair by students, which is another indicator that the system helped improve question quality. On the other hand, the use of QuesGen was not associated with a distribution of DI scores that would indicate that use of the system improved the quality of the question's distractors. Furthermore, QuesGen was not effective at enticing instructors to use the question quality checklist in a way that would allow them to avoid making technical errors in their questions. The checklist functionality seems to have the right idea, but its implementation in QuesGen was flawed.

In the next chapter, the contributions of this work and the logical next steps that follow from this discussion will be presented.

	i3	i4	i5	i6	i7	i8	i9	i10	i11	i12	i13	i14	i15	i16	i17	i18	i19	i20	i21	i22	i23	
i3	1.000																					
i4	0.085	1.000																				
i5	-0.069	-0.089	1.000																			
i6	0.188†	0.497§	-0.300§	1.000																		
i7	0.683§	0.076	-0.092	0.166†	1.000																	
i8	0.143†	0.177†	-0.044	0.191§	0.139†	1.000																
i9	0.086	0.061	-0.064	0.053	0.032	0.499§	1.000															
i10	0.012	0.187†	0.079	0.094	0.015	0.145†	0.056	1.000														
i11	0.304§	0.200§	-0.069	0.228§	0.433§	0.094	-0.028	0.150†	1.000													
i12	0.295§	0.130†	-0.068	0.188†	0.505§	0.105*	-0.011	0.137†	0.438§	1.000												
i13	0.108*	0.172†	-0.166†	0.215§	0.216§	0.041	0.053	-0.003	0.233§	0.2079§	1.000											
i14	0.085	0.133†	0.104*	0.022	0.066	0.102*	0.054	0.020	0.049	0.073	0.038	1.000										
i15	0.168†	0.107*	-0.044	0.056	0.148†	0.333§	0.218§	0.035	0.111*	0.136†	0.002	0.067	1.000									
i16	0.093	0.121*	-0.078	0.192§	0.232§	0.049	0.034	0.067	0.038	0.219§	0.179†	-0.031	0.017	1.000								
i17	0.038	-0.061	-0.010	0.084	0.010	0.049	-0.025	-0.049	-0.039	0.093	0.063	0.047	0.053	-0.015	1.000							
i18	0.063	0.076	0.018	0.118*	0.030	0.590§	0.428§	0.061	-0.037	0.007	-0.021	0.011	0.356§	0.008	-0.012	1.000						
i19	-0.049	0.020	0.114*	-0.024	-0.059	-0.024	0.019	0.024	-0.044	-0.052	-0.013	-0.001	0.001	-0.017	-0.039	0.014	1.000					
i20	-0.023	0.058	0.058	0.096	-0.028	0.060	0.082	-0.027	-0.021	-0.025	-0.029	0.041	0.088	-0.008	-0.019	0.141†	0.103*	1.000				
i21	-0.179†	-0.191§	0.400§	-0.467§	-0.204§	-0.217§	-0.072	-0.010	-0.237§	-0.197§	-0.229§	0.027	-0.070	-0.079	-0.041	-0.127*	0.072	-0.005	1.000			
i22	0.129†	0.055	-0.067	-0.031	0.080	-0.055	0.033	0.022	0.047	-0.041	0.072	0.057	0.003	-0.014	-0.031	-0.005	0.040	-0.017	-0.018	1.000		
i23	-0.027	0.003	-0.108*	0.013	-0.032	0.004	0.029	-0.030	-0.024	-0.028	-0.033	0.064	0.098*	-0.009	-0.021	0.059	-0.088	-0.011	0.063	0.256§	1	

* $p < 0.05$, † $p < 0.01$, ‡ $p < 0.001$, § $p < 0.0001$

Table 10.5 Correlation Matrix for Items on the Item-review Instrument

CHAPTER 11

CONCLUSION AND FUTURE WORK

11.1 The Context of This Dissertation

This thesis began as an attempt for one overwhelmed instructor to develop some software that would help him become more efficient. In the process, the goal grew and the research became finding a way to build systems that can help teachers in general become both more efficient and more effective. The first step in this direction was to develop a tool that could help teachers get better at one small part of their teaching practice—the development of high-quality multiple-choice questions. At one point along this journey, the researcher joked that QuesGen had an ulterior motive. He said that the real goal of QuesGen was to communicate to teachers that the development of MCQs was so difficult that they would abandon their use altogether. QuesGen was not successful in a clear and unequivocal way in achieving this goal, nor in achieving its stated goal, which was to help teachers get better at writing MCQs. However, QuesGen did have some small successes, and the study of QuesGen revealed a great deal of information that lights the way for future work in this area. This final chapter describes the concrete contributions that this thesis has made and lays out the work that is yet to be done.

11.2 Contributions

11.2.1 New, Web-based Software Tools for Writing and Evaluating MCQs

The first and most obvious contribution of this work is the QuesGen system itself. QuesGen was used successfully to write over 200 questions and deliver web-based video lectures, quizzes, and questionnaires to over 800 students. At the same time, QuesGen served as a massive data-collection engine. The results of the students' quizzes were used to dynamically create a personalized follow-up questionnaire for each student. The system randomized the order of the questions and presented the expert judges with item-review instrument questionnaire for each of the 200 questions that were evaluated. The system calculated the discrimination index, tallied the results of the student questionnaires, and judges' evaluations and compiled detailed question analysis reports that were then distributed to the participating instructors. Not only did several of the instructors, but the expert judges asked if this system would be available for them to use in the future. Clearly, QuesGen as a suite of tools has value for teaching, for learning, for research, and for the evaluation of questions.

11.2.2 A Clearer Picture of the Dimensions of MCQ Quality

The second contribution of this work is a clearer picture of the dimensions of MCQ quality. Even having read the literature, prior to doing this work it was unclear to this researcher how to interpret such statistics as the discrimination index, and how to use DI to inform the writing of future questions. It is painfully clear now that this research is completed that it is very important for instructors and students to carefully analyze the results of multiple-choice tests. It is all too easy to copy scores from a Scantron sheet

into a gradebook without ever going back to find out if maybe the reason so many people missed question #17 was because it was mis-keyed. It is also clearer that not only is it possible for MCQs to address higher-order thinking skills, the questions do not really provide much information about what students understand unless they address higher-order skills. It is also clear that all of the empirically validated technical question flaws are things that are relatively straightforward for most any teacher to understand and avoid. All that remains is to find the interface that is best able to help them learn how avoid these errors, which brings us to the third contribution.

11.2.3 A New Model for Teacher Professional Development Software

QuesGen was not an evaluation of the software model described in Chapter 4, but it was the first version of a tool that can eventually lead to the evaluation of this model. The ideas in the transtheoretical model of behavioral change are general enough to be applied to the task of inculcating best practices in most any field of human endeavor. Combined with the domain-specific insights of the KBA framework into the reasons behind effective and lasting teacher professional development the opportunity arises for a class of software systems that could greatly improve our educational systems in a cost effective and distributed fashion. It is the goal of this researcher to pursue this vision in the years to come, which brings us to the discussion of future work.

11.2.4 Data-Gathering Integrated Into System Design

A great deal of the data collection for the evaluation of QuesGen was done by the system itself as the data-gathering functionality was included in the implementation. These analytical tools were not originally part of the design of the system, but were added to

ease the process of evaluation. However, it turned out that this functionality proved to be useful not only to the researcher, but also to the participants in the study who were very interested to see the data that were produced and reflect upon their own practice using that data. In retrospect it is apparent that such tools fit into the transtheoretical model under the heading of self-reevaluation and should therefore be included explicitly in future versions of QuesGen. Furthermore, it seems intuitive that providing analytical tools based upon system usage is a generalizable strategy that could be employed when building most any type of system, particularly one designed as experimental or for research.

11.2.5 Instruments for Understanding Question Quality

The student survey and the item-review instruments provided insight into the quality of the questions that were developed. Instructors clearly indicated that they would like to get more feedback from their students with respect to the quality of the questions that they have asked. They were particularly interested in the open-ended feedback that students provided about the questions. The item-review instrument, while not providing the resolution that the researcher had hoped for, was nonetheless useful in focusing attention on various aspects of the technical quality of questions. It will certainly serve as the basis for future instruments to be designed in this effort.

11.3 Future Work

This thesis suggests future work in two areas: development of the next version of QuesGen, and also development of some research questions that follow on from issues that arose during this project.

11.3.1 QGv2

The following is a list of implications and goals for the next version of QuesGen that follow on from the discussion of the results above. The use of a participatory design strategy for incorporating these features is strongly indicated.

- **More obtrusive features**

Perhaps “obtrusive” is not the correct word, but a major problem with the system was that the teachers did not use the features that were designed to help them. Given the history of IS implementations, this is not really a surprising result, but it is disappointing nonetheless and future versions of the tool need to work harder to convince users to take advantage of what it has to offer. In particular, the question templates, given their association with higher-order questions, should be a focus.

- **“Smarter” incorporation of the checklist**

While it is clear that the question quality checklist has lots of useful suggestions for avoiding flaws, not all of them apply to every question. A way needs to be found to prompt the user intelligently to check for flaws. This might mean the incorporation of grammar-checking software and/or natural language processing tools for flagging potential problems. Wisdom on how to go about this might come from the work of other researchers such as Joanna McGrenere’s work on “bloat” (e.g. McGrenere 2002).

- **More responsiveness**

While the usability studies done with the tool clearly indicated that it had to contain a WYSIWYG editor for writing questions and incorporating images into

questions, when implemented this greatly decreased the responsiveness of the system. The editor also introduced bugs which made the interface basically unusable to at least one instructor. In the months since QuesGen was implemented the state of the art of web-based applications has advanced tremendously, and solutions to the problems encountered here have been developed.

- **Explicit incorporation of the student surveys and QARs**

When presented with their students' feedback on the question analysis reports, nearly all of the participant instructors said that they would really like to be able to get this type of feedback from their students on a regular basis. Originally it was included in QuesGen as a way to generate data for the thesis, but it seems that it should become a feature in future versions of the tool.

- **The ability for teachers to write and include their own objectives**

Actually, this functionality already exists, but was disabled for the purposes of the study because it was perceived to be another source of variance that needed to be controlled in the study.

- **More/Better incorporation into existing systems, i.e. Blackboard or WebCT**

QuesGen does have the ability to export questions in the text format needed for both Blackboard and WebCT, but with the new APIs being released by these companies, even tighter integration is possible. It may be possible to replace or supplant the existing tools in these systems with higher-quality ones.

- **Exploring the incorporation of concept inventories**

Concept inventories (e.g. Hestenes 1992) have been incorporated into physics education and are designed to test the understanding of theoretical knowledge as opposed to computational knowledge. This focus bears a strong correlation to the work of QuesGen, which seeks to get instructors to ask questions that assess understanding of higher-order levels of knowledge.

11.3.2 Research

In addition to the new and improved functionality of a new version of QuesGen, a number of interesting research questions follow on from the work that has been done so far.

The first question is what is the long-term impact of a tool like QuesGen? Relatively modest gains in question quality were seen in limited areas with the current study in which the tool was used only once. It seems likely from the evidence that QuesGen will not begin to have a really strong impact until it has been used over a longer period of time, say an entire semester, or perhaps an entire academic year. When a significant number of the currently existing problems with the system have been fixed, the next study to be run should be longitudinal.

A second question that arises from this study is what differences in question quality gains can be attributed to experience and what differences can be attributed to instructor experience. Since these two variables were confounded in this study, future studies should take more care to separate these.

Following on from the previous question, if instructor experience has such a strong impact, does QuesGen speed up the natural learning processes that teachers go through as they learn to use MCQs as part of their assessment routines.

A fourth question to be asked is how would have this research have been different if more care had been spent on creating truly high-quality educational objectives to feed to the teachers. In the K12 environment, most states have very tightly prescribed educational objectives at every grade level. In this study, the quality of the objectives seems to have had an impact on the quality of the questions written by the teachers. Does it in turn have an impact on the quality of their teaching? What if instructors are allowed to generate their own questions?

A fifth question is what role could QuesGen play in educating graduate student teaching assistants in assessment? While it appears that instructors improve over time in their ability to write high-quality MCQs, one of the strengths of QuesGen may be to catalyze that professional development. If QuesGen were successful in helping novice instructors reach proficiency faster, then it would certainly be a valuable addition to many graduate programs.

In addition to these questions, there are some methodological issues that could be addressed. One idea would be to re-run the experiment with less control over some of the variables like instructor experience and course content. It is possible that the lack of variance in the results was caused by too much control, and therefore a loosening of the controls might make for a more powerful test of QuesGen's capabilities. It would also be desirable to run a within-subjects test to gauge QuesGen's ability to foster improvement within a single instructor. Another issue is the consideration of the role of quizzes within

an instructor's pedagogy. In this study, a clear bias was exhibited towards questions that tap higher-order thinking, but there are defensible arguments as to why an instructor would want to ask questions that focus on memory at different points within the course of instruction. A follow-up study would take these motivations more into consideration and adjust the evaluation of question quality to accommodate the different uses for quizzes.

These are just some of the questions that could be pursued. There are certainly many others. Ultimately, the goal is to move in the direction of building a system that would allow the full system model based on TTM and KBA to be tested.

11.4 Conclusion

There is a long way to go in the development of systems that can serve as professional development tools for teachers. QuesGen has taken some first steps in this direction. In the process, a great deal was learned about question quality, and about building systems to help teachers write questions that are of higher quality. Thank you for reading this dissertation.

APPENDIX A

STUDENT CONSENT TO PARTICIPATE IN RESEARCH PROJECT

Identification of Investigators & Purpose of Study

You are being asked to participate in a research study conducted by Morgan C. Benton from James Madison University, Integrated Science and Technology Department. The purpose of this study is to determine the effectiveness of QuesGen, a web-based tool designed to help instructors write better multiple-choice questions. Your part in this study will be evaluating the questions written with this tool.

Potential Risks & Benefits

The investigator does not perceive more than minimal risks from your involvement in this study.

Potential benefits from participation in this study include:

1. Contributing to research that aims to improve the relevance and effectiveness of instruction at JMU.
2. An opportunity to provide anonymous feedback to your instructor on the quality of the questions on the quiz you just took.
3. A chance to win one of three \$25 gift certificates to Barnes & Noble.

Research Procedures

This study consists of an online survey and will take approximately 10 minutes. Recently you were asked to watch an online lecture and complete a quiz assessing your understanding of the concepts introduced in that lecture. As a participant in this survey, you will be shown each of the questions that you encountered on that quiz and asked to

rate the difficulty, fairness, and clarity of each question, as well as possibly indicating why you think you missed the question (if you missed the question). Your responses will be used, in part, to help determine the quality of the questions. Your feedback will also be shared (anonymously) with your instructor to help him/her more fully understand students' perception of the quiz questions they write.

The data that will be collected in this study will be your responses to the quiz questions, and your feedback on the quality of the quiz questions. Your feedback will be anonymous, i.e. the researcher will never associate your name with your comments. However, there is a small chance that your instructor might guess which students made which comments. Therefore, it is important that you not make comments that would be personally identifiable. Your responses on the survey, and your choice to participate or not, will have absolutely no effect on your semester grade.

Should you decide to participate in this confidential research you may access the survey by following the web link located under the "Giving of Consent" section.

Confidentiality

If you decide to participate in this study, your identity will remain confidential. Only the researcher will have the ability to associate your name with your responses on the survey. At no time will your name or identity be made public in connection with your responses on this survey. The results of this project will be coded in such a way that the respondent's identity will not be attached to the final form of this study. The researcher retains the right to use and publish non-identifiable data. It is anticipated that the results of the study will be published as part of the researcher's Ph.D. thesis, and also at appropriate conferences or research journals. Aggregate data will be presented

representing averages or generalizations about the responses as a whole. All data will be stored in a secure location accessible only to the researcher. Upon completion of the study, all information that relates individual respondents to their answers will be destroyed. Final aggregate results will be made available to participants on the researcher's website at the conclusion of the study. The link to this website will be provided at the conclusion of the survey. We will also be asking you to list an email address by which you can be contacted. This will be used both to mail you the results of this study (if requested) and to notify you if you have won any of the drawings for gift certificates.

Participation & Withdrawal

Your participation is entirely voluntary. You are free to choose not to participate. Should you choose to participate, you can withdraw at any time without consequences of any kind. However, once your responses have been submitted and anonymously recorded you will not be able to withdraw this information from the study. If you do withdraw from the study, your name will not be included in the drawing for prizes.

Questions

You may have questions or concerns during the time of your participation in this study, or after its completion. If you have any questions about the study, contact:

*Morgan C. Benton
ISAT Department, MSC 4310
James Madison University
Harrisonburg, VA 22807
bentonmc@jmu.edu
(540) 568-6876*

Giving of Consent

I have read this consent form and I understand what is being requested of me as a participant in this study. I freely consent to participate. The investigator provided me with a copy of this form via his website. I certify that I am at least 18 years of age. By clicking on the link below, and completing and submitting this confidential survey, I am consenting to participate in this research.

<https://www.quesgen.net>

Morgan C. Benton

Name of Participant (printed)

Name of Researcher (Printed)

Name of Participant (signed)

Name of Researcher (Signed)

Date

Date

For questions about your rights as a research subject, you may contact the chair of JMU's Institutional Review Board (IRB). Dr. David Cockley, (540) 568-2834, cocklede@jmu.edu.

APPENDIX B

INSTRUCTOR CONSENT TO PARTICIPATE IN CONFIDENTIAL RESEARCH

Identification of Investigators & Purpose of Study

You are being asked to participate in a research study conducted by Morgan C. Benton from James Madison University, Integrated Science and Technology Department. The purpose of this study is to determine the effectiveness of QuesGen, a web-based tool designed to help instructors write better multiple-choice questions.

Potential Risks & Benefits

The investigator does not perceive more than minimal risks from your involvement in this study. Potential discomforts involve those one would normally associate with learning how to use a new software tool.

Potential benefits from participation in this study include:

1. The chance to use a new form of instructional technology.
2. An increased understanding of how to write multiple-choice questions.
3. Questions you can use to assess your students' mastery of course content.
4. Feedback from students and expert judges on the quality of your questions.
5. The chance to help a colleague complete his dissertation research.

As a token of appreciation for your participation you may select one of the following: 1) a personalized, frangible piece of Japanese calligraphy, or 2) a gourmet, home-cooked Japanese meal. Both of these will be provided by the investigator's wife, who is Japanese and a professional calligrapher and cook.

Research Procedures

You have been selected for participation in this study because you are one of several teachers that teach a large, introductory course in JMU's Gen Ed program. All of the instructors that teach the same course as you are also being asked to participate. The success of this study depends upon having as close to 100% instructor participation as possible. The duration of the study will be about two weeks. Your personal time commitment during this period is estimated to be approximately 4-6 hours. As a participant, you will be asked to do the following:

1. Watch an online video lecture designed to be a single unit of instruction in the course that you teach.
2. Use the online software tool, QuesGen, to write 10 multiple-choice questions that will assess students' mastery of the content in the video lecture. The system will log your usage.
3. Oversee and administer your class as they watch the video lecture and respond to the 10 questions.
4. Solicit your students' participation in a follow-up questionnaire after the quiz.
5. Respond to a short questionnaire about your own experiences with QuesGen.
6. Participate in a follow-up video-taped interview session that will explain the goals of QuesGen and the study in more depth and solicit qualitative feedback on your experiences.

You will be asked to refrain from discussing your experiences with QuesGen with your colleagues during the study, but will have the opportunity to do after the study has completed.

Confidentiality

Because of the nature of the study, your participation will not be kept secret from your colleagues who are also participants in the study. However, your identity will be kept confidential in any published works or presentations that result from this research. Furthermore, any questions that you develop, or any responses that you make on the questionnaires will not be shared with your colleagues in any identifiable way. The video tapes of the interviews will be kept in a locked filing cabinet in the researcher's office, to which no one will have access but the researcher. The tapes will be destroyed upon the completion of their analysis. Upon request, you may be present to witness the destruction of the tape of your interview.

Participation & Withdrawal

Your participation is entirely voluntary. You are free to choose not to participate. Should you choose to participate, you can withdraw at any time without consequences of any kind.

Questions

You may have questions or concerns during the time of your participation in this study, or after its completion. If you have any questions about the study, contact:

*Morgan C. Benton
ISAT Department, MSC 4310
James Madison University
Harrisonburg, VA 22807
bentonmc@jmu.edu
(540) 568-6876*

Giving of Consent

I have read this consent form and I understand what is being requested of me as a participant in this study. I freely consent to participate. The investigator provided me with a copy of this form.

Morgan C. Benton

Name of Participant (printed)_____
Name of Researcher (Printed)_____
Name of Participant (signed)_____
Name of Researcher (Signed)_____
Date_____
Date

For questions about your rights as a research subject, you may contact the chair of JMU's Institutional Review Board (IRB). Dr. David Cockley, (540) 568-2834, cocklede@jmu.edu.

APPENDIX C

ITEM REVIEW INSTRUMENT USED BY EXPERT PANEL

Which of the following is the best example of [target concept]?

- A. [example 1]
- B. [example 2]
- C. [example 3] (*key*)
- D. [example 4]

Section 1: Educational Objectives

Below are the educational objectives that were explicitly stated in the lecture slides for this lecture. Please rate the degree to which the question in the box above assesses mastery of each objective.

Objective 1: Students will be able to ...

1. The question in the box above assesses students' mastery of this objective:
Not at all Completely

Objective 2: Students will be able to ...

2. The question in the box above assesses students' mastery of this objective:
Not at all Completely

Objective 3: Students will be able to ...

3. The question in the box above assesses students' mastery of this objective:
Not at all Completely

[... repeated for all objectives for the unit of instruction]

Section 2: Bloom Classification

4. To which Bloom category should the question in the box above be assigned?
- Recall
 - Comprehension
 - Application
 - Analysis
 - Synthesis
 - Evaluation
5. Given the content in the video lecture, could this question be answered correctly solely by remembering the content of the lecture?
- Yes
 - No

Section 3: Technical Flaws

For each of the following technical flaws, please rate the degree to which you agree with the following statements.

6. The answer marked as the key above is the best answer.
Strongly Disagree Strongly Agree
7. This question clearly addresses a stated educational objective.
Strongly Disagree Strongly Agree
8. The concept examples used in this question are novel.
Strongly Disagree Strongly Agree
9. This question clearly addresses exactly one educational objective.
Strongly Disagree Strongly Agree
10. This question has exactly one correct answer.
Strongly Disagree Strongly Agree
11. All of the answer choices for this question are plausible.
Strongly Disagree Strongly Agree
12. This question's distractors are based upon likely student misconceptions.
Strongly Disagree Strongly Agree
13. This content assessed in this question is trivial.
Strongly Disagree Strongly Agree
14. The content in this question is based upon opinion.
Strongly Disagree Strongly Agree
15. This is a trick question.
Strongly Disagree Strongly Agree
16. This question depends upon cultural knowledge.
Strongly Disagree Strongly Agree
17. This question is grammatically correct.
Strongly Disagree Strongly Agree
18. The answer choices in this question use parallel sentence structure.
Strongly Disagree Strongly Agree
19. The vocabulary in this question is appropriate to the student level.
Strongly Disagree Strongly Agree

20. This question is concise.
Strongly Disagree Strongly Agree
21. The answer to this question is not obvious.
Strongly Disagree Strongly Agree
22. This question follows best practices in the use of the words *NOT* and *EXCEPT*.
Strongly Disagree Strongly Agree
23. This question inappropriately uses humor.
Strongly Disagree Strongly Agree
24. Students can answer this question correctly purely from memory.
Strongly Disagree Strongly Agree
25. This question avoids using “all” or “none of the above.”
Strongly Disagree Strongly Agree
26. This question avoids using absolutes such as “every” or “never.”
Strongly Disagree Strongly Agree

APPENDIX D

QUESGEN: QUESTION ANALYSIS REPORT (EXAMPLE)

Title: Power Form- Question 3

Objective: Define and identify the 3 forms of power.

Template: Troubleshooting

Question: Diagnose the problem in the following scenario. Joey tells his roommate John, "Clean up this messy room or else." John says "Sure" but proceeds to avoid cleaning up his room even though he had planned on cleaning it up before speaking to Joey. In fact, he makes his room messier on purpose after their conversation.

- 65% (15/23) A. *Joey is operating from a dominance perspective and John is operating from a prevention perspective.*
- 17% (4/23) B. John is operating from a dominance perspective and Joey is operating from a prevention perspective.
- 17% (4/23) C. Joey and John are both operating from a dominance perspective.
- 0% (0/23) D. Joey and John are both operating from a prevention perspective.

Students' Evaluation of the Question

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
The question was difficult	13% (2/16)	56% (9/16)	19% (3/16)	13% (2/16)	0% (0/16)
The question was fair	0% (0/16)	0% (0/16)	25% (4/16)	38% (6/16)	38% (6/16)
The wording was clear	6% (1/16)	13% (2/16)	25% (4/16)	25% (4/16)	31% (5/16)

Words students said they didn't know:

Making it messier made it seem like he wanted dominance to prove a point, prevention perspective

Why students said they missed the question:

The question was confusing: 25% (1/4)

My interpretation was different than the instructor's: 50% (2/4)

Careless error: 25% (1/4)

Item-review Panel Evaluation of the Question

The panel rated the degree to which this question assessed understanding of the following educational objectives, which were taken from the lecture slides:

Objective #1: Define power and give an original example of power in their own lives.

Panelist #1: not at all

--	--	--	--	--

 completely

Panelist #2: not at all

--	--	--	--	--

 completely

Objective #2: Define and identify the 3 forms of power.

Panelist #1: not at all

--	--	--	--	--

 completely
 Panelist #2: not at all

--	--	--	--	--

 completely

Objective #3: Define and identify the 5 types of power resources.

Panelist #1: not at all

--	--	--	--	--

 completely
 Panelist #2: not at all

--	--	--	--	--

 completely

Objective #4: Use the terms to label which power forms and resources are present when presented within a communication scenario.

Panelist #1: not at all

--	--	--	--	--

 completely
 Panelist #2: not at all

--	--	--	--	--

 completely

The panel agreed that this question adhered to the following best practices:

- The answer marked as the key above is the best answer.
- This question clearly addresses a stated educational objective.
- The concept examples used in this question are novel.
- This question clearly addresses exactly one educational objective.
- This question has exactly one correct answer.
- All of the answer choices for this question are plausible.
- This question's distractors are based upon likely student misconceptions.
- The question addresses salient knowledge.
- The content in this question is NOT based upon opinion.
- This is a NOT trick question.
- This question does NOT depend upon cultural knowledge.
- This question is grammatically correct.
- The vocabulary in this question is appropriate to the student level.
- The answer to this question is not obvious.
- This question follows best practices in the use of the words NOT and EXCEPT.
- This question uses humor appropriately.
- Students can NOT answer this question correctly purely from memory.
- This question avoids using "all" or "none of the above."
- This question avoids using absolutes such as "every" or "never."

The panel agreed that this question violated the following best practices:

- The panel didn't agree that your question violated any of the best practices.

Level of Bloom's Taxonomy targeted by this question:

- Panelist #1: application
- Panelist #2: analysis

Discrimination Index: 0.333333333333

APPENDIX E

FOLLOW-UP INTERVIEW GUIDE

Introduction: First of all, thank you for participating in this study. Your time and effort are greatly valued and appreciated. The purpose of this interview is to capture your thoughts, attitudes, and feelings towards multiple-choice questions, towards learning and assessment, towards educational objectives, and towards various design features in QuesGen. You have already spent time using QuesGen to develop a quiz and deliver that to your students. During this interview I will spend time with you discussing your experiences with QuesGen. We will go over several of the questions that you wrote, and I will ask you about the process you went through to generate those questions. I will show you several measures of how your questions performed on the quiz—not only your students’ answers, but also their subjective feedback on the questions, as well as the evaluation of an item-review panel that was convened to rate the quality of the questions—and I will ask for your reactions to these evaluations.

I expect this interview to take approximately 45 minutes. I remind you that you are free to end the interview at any time and for any reason. You are also free to ask me to turn off the videotaping at any time. I am the only person who will have access to the video recording of this interview. While I may use quotations or paraphrases of what you say in this interview for the purposes of reporting on the results of this experiment, these quotations will never be made in a way that is personally identifiable. Your identity will remain confidential. I will destroy the actual recording of this interview once this study has been completed. Please feel free to ask questions at any time. Do you have any questions now about this process?

Background Questions

1. What is/are your field or fields of expertise?
2. How long have you been a university teacher?
3. What is the course you teach that is part of this study?
4. What other courses do you teach, or have you taught regularly in the last five years?
5. Have you ever taught anywhere besides JMU? If so, where? For how long did you teach there?
6. How would you characterize JMU students (in relation to students at other places you have taught—if applicable)?

Questions about the course and instructor's attitudes toward students in the course

7. How many times have you taught this course?
8. In your own words, could you briefly describe the main objective of this course and how it fits into the curriculum?
9. Generally speaking, how would you describe or characterize the students who are taking this course? Probes: number of students, year in college, ages, genders, attitude toward the course, ability level, level of preparation, motivation to study.
10. How successful are students at mastering the objectives of this course?
11. What is the most difficult concept or part of the course to teach, and why?

Questions about the unit of material used for the study

12. Did you feel that the video lecture was at the level of the students, below the level, above the level? Why?
13. Did you find the lecture to be boring, okay, interesting? Why?
14. Did you feel the lecture was very easy, about right, very hard for the students to understand? Why?
15. [Show a list of the objectives from the lecture slides.] Do you think that the educational objectives listed at the beginning of the lecture were the right ones for this unit? Would you have added, removed or changed any of them? If so, which ones and how?

Questions about attitudes toward MCQs

16. Do you normally write multiple choice questions for exams? If so, or if not, why?
17. Will the multiple choice question practice you had with the video lecture encourage you to use multiple choice questions more often in exams? Why or why not?
18. Which do you prefer to write? Multiple choice questions or essay questions? Why?
19. Has this preference changed in any way after you used Quesgen? If so, in what way?
20. Do you think multiple choice questions are a fair way to assess student's performance? Why or why not?
21. Has your opinion of the fairness of multiple choice questions changed after creating them for the video lecture? If so, in what way has it changed?
22. Do you think multiple choice questions are easy to write? Why or why not?
23. Do you think the practice you had with the video lecture made them easier to write? Why or why not?
24. Under what circumstances do you think multiple choice questions are appropriate to use for student assessment?
25. Under what circumstances do you think that multiple choice questions should not be used for student assessment?
26. Have you changed your opinion on when to use and not use multiple choice questions after the practice you had in writing multiple choice questions for the video lecture? Please explain why or why not.

27. How good a multiple choice question writer have you been in the past? If you have never written multiple choice questions but only selected them from a test bank, then indicate how good you think your skill was in selecting good questions from such a test bank.
28. Did the practice you had in writing multiple choice questions for the video lecture make you a better multiple choice question writer? In what ways do you think it made you better? In what ways did the practice not have any effect?

Questions about the impact of specific aspects of QuesGen

29. Have you used teaching objectives for each of your lectures in your courses? If so, can you explain how you used them?
30. Have you used teaching objectives to generate tests for your students? If so, explain how you used them?
31. Did your practice in writing multiple choice questions for the video lecture make you more aware of using teaching objectives to design questions? If so, can you give an estimate of how much of an impact this practice will have on your future question writing behavior (none at all to ...a significant amount)
32. Before you practiced writing questions for the video lecture, were you aware of basic dos and don'ts for writing multiple choice questions (e.g., do not use the word "not")
33. Would you now like a written set of these rules around for future question writing?
34. Before you practiced writing questions for the video lecture, were you aware of levels of learning that could be assessed through different templates? If so, how much were you aware of this? Can you describe your depth of knowledge?
35. Would you now like to have these templates available for future question writing?
36. Of the three supports given you for multiple choice question writing (teaching objectives, dos and don'ts, templates) which one do you think is the most important? Why?
37. Of the three supports given you for multiple choice questions writing, which one do you think is the second most important? Why?
38. Please explain why the final support is the least important.
39. Do you think that having these supports made your questions better? Why or why not?
40. Do you think that student assessment is more accurate if questions are written with these support tools? Explain your answer.
41. Do you think student assessment is fairer if questions are written with these support tools? Explain your answer.

APPENDIX F

STUDENT QUESTIONNAIRE

Answer the following questions about question #1 shown in the box below:

1. Which of the following is a conclusion that can be made after reading Borning's (1987) article on computers and nuclear war?

- A. Using Tomahawk missiles reduces the chances for nuclear war
- B. National Missile Defense (NMD) makes all of Borning's conclusions obsolete
- correct** → C. The complexity of nuclear weapons systems demands extra special attention from us as developers since we are literally playing with fire
- your answer** → D. Russia is no longer a threat to the United States in terms of nuclear capability

8. This question was difficult.
Strongly Disagree Strongly Agree
9. This was a fair question.
Strongly Disagree Strongly Agree
10. The wording of this question was clear.
Strongly Disagree Strongly Agree
11. How many words in this question and answer choices did you not know or understand?
- a. Zero
 - b. One or two
 - c. Three or more
12. If you answered "b" or "c" in the previous question, which words gave you trouble?
- _____
13. If you missed this question, why do you think you did? Please select the reason that best matches your situation.
- a. The question was too hard
 - b. The question was confusing
 - c. My interpretation was different than the instructor's
 - d. I didn't study/prepare enough beforehand
 - e. Careless error
 - f. Other (please specify) _____

APPENDIX G

INTRODUCTORY EMAIL SENT TO PARTICIPATING INSTRUCTORS

Dear Participant,

Thank you for agreeing to participate in my study! It starts today. Please read this email CAREFULLY. You may want to print it out.

In this email:

- Your instructions
- Other things to do and NOT to do
- Detailed schedule
- FAQ

Your Instructions for THIS WEEK

1. Log in to <http://www.quesgen.net>
 1. Username: mcbenton
 2. Password: mcb17que
2. Read the consent form

Click the "I Consent" button at the bottom to agree.
3. Watch the QuesGen Tutorial Video (a little over 15 minutes)
4. Watch your colleague's video lecture on Power

It's about 21 minutes long. Download the slides if you wish.
5. Use QuesGen to write 10 multiple-choice questions

Due date: Noon, Thursday 2/22. The questions should assess students' mastery over the content in the video you just watched. Please stick to questions about the video lecture. Do NOT write questions that address other things you may have covered in your class.
6. Please tell your students about the online QuesGen activity
 1. Tell them in class
 2. Give them the handouts

Handouts will be in your box by the end of today. Please give them out THIS WEEK.
 3. Tell them to expect email with detailed instructions next Monday morning, 2/26
7. Please email me when you finish your questions

When I hear from you I will unlock the follow-up survey for you.
8. Please complete the follow-up survey (takes about 15 min.)

Other things to do and NOT to do

- Please finish your questions by NOON, THURSDAY 2/22, three days from now.
- Please expect to hear from me daily with friendly reminders, or status checks.
- Please do NOT talk about the study with your colleagues or students until AFTER everyone has finished their questions and follow-up survey. This will probably be by the end of next week.

- Please CALL ME AT ANY TIME (973) 495-7736, if you have questions or problems. I am on call 24/7 until this thing is done. My future is riding on this going well.
- Please send me your course roster(s) with students' names and email addresses.

Detailed Schedule

- **Today, Monday 2/19: Study Begins**
Teachers will log into QuesGen, watch the videos, and write questions. Teachers will tell students about next week's study and distribute handouts.
- **Noon, Thursday, 2/22: Teachers Complete Questions**
Questions will be written and loaded into QuesGen for quiz delivery the following Monday.
- **9AM Monday, 2/26: Students Receive Email Instructions**
Students will get a detailed email (like this one, but shorter) explaining exactly what they are to do. They will watch the online video lecture, take the quiz, and then a follow-up survey.
- **Thursday, 3/1: Students complete quiz and survey**
Teachers will be able to access quiz grades.
- **Saturday, 3/3: Expert judges will evaluate question quality**
An expert item-review panel from CARS will do an anonymous review of the quality of the questions written by the teachers.
- **Monday, 3/5-Friday, 3/16: Follow-up Interviews**
I will contact you to schedule a 1-on-1 debriefing session. This should last about 45 minutes, and I will go over the results of your question analysis with you.
- **Saturday, 3/17-Friday, 4/13: Data analysis**
Draft thesis will be delivered to my committee on 4/13. Defense is scheduled for 5/11. I will provide a summary of the results for everyone by the end of the semester.

FAQ

1. How will I/my students get their grades?
You will be able to log into QuesGen and download them. Students will be able to see their scores and exactly how they answered each question.
2. What will the students see when they log in?
You can watch a short video of what the student interface looks like.
3. How do I get the questions into BlackBoard?
For the purposes of this study, that is not necessary because students will take the quiz right in QuesGen. After the study is over, when you log into QuesGen there will be a link that will allow you to download your questions in a format that can then be imported into BlackBoard.
4. How do I know if my students are cheating or not? Is this system secure?
You won't, and it isn't. A very low-stakes quiz was chosen on purpose for this study. The study has been crafted to minimize students' motivation to cheat. The results of the study are not particularly sensitive to cheating.

5. After this is over, will I still be able to use QuesGen?
Maybe not right away. The version of QuesGen built for this study has serious scalability limitations. After this research is complete, further development of the system is planned.

Again, thank you. If you have any questions whatsoever, please do not hesitate to contact me.

Sincerely,

Morgan Benton

Assistant Professor
ISAT Department, MSC 4310
James Madison University
Harrisonburg, VA 22807
bentonmc@jmu.edu
(540) 568-6876 (office)
(973) 495-7736 (mobile)

APPENDIX H

INTRODUCTORY EMAIL SENT TO STUDENT PARTICIPANTS

Dear Student,

This week [your class] students will be participating in an online learning experience. You will watch an online lecture (about 12 minutes), take a quiz on the content of that lecture (about 10 minutes), and then be asked to complete a questionnaire (another 10 minutes). Completing the questionnaire is optional, but if you do, you will have your name entered in a drawing for one of three \$25 gift certificates to Barnes & Noble, iTunes Music Store, or Amazon.com--your choice!

Here Are Your Instructions--Due Date: 9AM, 3/1/2007

1. Log in to <http://www.quesgen.net>
 1. Username: stud
 2. Password: letmein
2. Watch a Video Lecture on Power & Communication
This last about 12 minutes.
3. Take a 10 Question Quiz
This will assess your understanding of the content of the video you just watched.
4. (Optional) Click the Link to Take the Follow-Up Questionnaire
5. Read the consent form
Click the "I Consent" button at the bottom to agree.
6. Respond to the Questionnaire
This should take about 10-15 minutes, tops.
7. That's it. You're done.

Please Note!

Approximately 1800 students will be accessing the videos and the quiz this week, so there may be times when the server is slow. If this happens please be patient, and/or try again at another time.

Why am I being asked to do the questionnaire?

QuesGen is a web-based system designed to help instructors write better multiple-choice questions. It is currently being tested, and you are being asked to help with the testing. Your responses on the questionnaire will help to determine whether or not QuesGen is effective.

Who is Morgan Benton?

Professor Benton is a new faculty member in the Integrated Science and Technology Department, and the creator of QuesGen. Your GKIN 100 instructor has graciously offered to help him test his system.

Again, thank you. If you have any questions whatsoever, please do not hesitate to contact me.

Sincerely,

Morgan Benton

Assistant Professor
ISAT Department, MSC 4310
James Madison University
Harrisonburg, VA 22807
bentonmc@jmu.edu
(540) 568-6876 (office)
(973) 495-7736 (mobile)

APPENDIX I

UTAUT INSTRUMENT USED IN EVALUATING QUESGEN

Instructions: Please indicate your level of agreement with the following statements.

1. I find QuesGen to be useful in my job.
Strongly Disagree Strongly Agree
2. Using QuesGen enables me to write multiple-choice questions more quickly.
Strongly Disagree Strongly Agree
3. Using QuesGen increases my productivity.
Strongly Disagree Strongly Agree
4. If I use QuesGen, I will increase my chances of getting a raise.
Strongly Disagree Strongly Agree
5. Using QuesGen will reflect positively on my annual performance evaluation (FAR).
Strongly Disagree Strongly Agree
6. My interaction with QuesGen is clear and understandable.
Strongly Disagree Strongly Agree
7. It is easy for me to become skillful at using QuesGen.
Strongly Disagree Strongly Agree
8. I find QuesGen easy to use.
Strongly Disagree Strongly Agree
9. Learning to operate QuesGen is easy for me.
Strongly Disagree Strongly Agree
10. People who influence my behavior think that I should use QuesGen.
Strongly Disagree Strongly Agree
11. People who are important to me think that I should use QuesGen.
Strongly Disagree Strongly Agree
12. My department and/or college have been supportive in the use of QuesGen.
Strongly Disagree Strongly Agree
13. In general, JMU/college/department has supported the use of QuesGen.
Strongly Disagree Strongly Agree

REFERENCES

- Anderson, L. W., and Krathwohl, D. R., *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, Longman, New York, 2001.
- Bandura, A., "Self-efficacy Mechanism in Human Agency," *American Psychologist*, Volume 37, Number 2, 1982, pp. 122-47.
- Bangert-Drowns, R. L.; Kulik, J. A.; and Kulik, C.-L. C., "Effects of Frequent Classroom Testing," *Journal of Educational Research*, Volume 85, Number 2, 1991, pp. 89-99.
- Bejar, I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., and Revuelta, J., "A Feasibility Study of On-the-Fly Item Generation in Adaptive Testing," *Journal of Technology, Learning, and Assessment*, Volume 2, Number 3, 2003.
- Benton, M. C.; Tremaine, M. M.; and Scher, J. M., "Computer Aids for Designing Effective Multiple-Choice Questions," *Proceedings of the Americas Conference on Information Systems*, New York, New York, 2004, pp. 3059-3066.
- Berdichevsky, D., and Neuenschwander, E., "Toward an Ethics of Persuasive Technology," *Communications of the ACM*, Volume 42, Number 5, 1999, pp. 51-58.
- Black, P., Harrison, C., Lee, C., Marshall, B., and Wiliam, D., *Assessment for Learning: Putting it into practice*, Open University Press, Berkshire, England, 2003.
- Black, P., and Wiliam, D., "Inside the Black Box: Raising Standards Through Classroom Assessment," *Phi Delta Kappan*, Volume 80, Number 2, 1998a, pp. 139-149.
- Black, P., and Wiliam, D., "Assessment and Classroom Learning," *Assessment in Education: Principles, Policy and Practice*, Volume 5, Number 1, 1998b, pp. 7-75.
- Blessing, S., "The Cognitive Tutor™: Successful Application of Cognitive Science," *Carnegie Learning®*, 2003.
- Bloom, B. S., "The Two Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring," *Educational Researcher*, Volume 13, 1984, pp. 4-16.
- Bloom, B. S., Hastings, J. T., and Madaus, G. F. "Handbook on the Formative and Summative Evaluation of Student Learning," New York, NY, 1971.
- Bock, R. D., Wolfe, R., and Fisher, T. H. "A Review and Analysis of the Tennessee Value-Added Assessment System," (State of Tennessee, Comptroller of the Treasury, Office of Education Accountability), 1996, pp. 94.

- Borko, H., "Professional Development and Teacher Learning: Mapping the Terrain," *Educational Researcher*, Volume 33, Number 8, 2004, pp. 3-15.
- Brookhart, S., "Teachers' grading practices: Meaning and Values," *Journal of Educational Measurement*, Volume 30, 1993, pp. 123-142.
- Brophy, J. E., "Research on the Self-Fulfilling Prophecy and Teacher Expectations," *Journal of Educational Psychology*, Volume 75, 1983, pp. 631-661.
- Butler, R., "Enhancing and undermining Intrinsic Motivation: The Effects of Task-involving and Ego-involving Evaluation on Interest and Performance," *British Journal of Educational Psychology*, Volume 58, Number 1, 1988, pp. 1-14.
- Campbell, R. J., Kyriakides, L., Muijs, R. D., and Robinson, W., "Differential Teacher Effectiveness: towards a model for research and teacher appraisal," *Oxford Review of Education*, Volume 29, Number 3, 2003, pp. 347-362.
- Campbell, R. J., Kyriakides, L., Muijs, R. D., and Robinson, W., "Effective teaching and values: some implications for research and teacher appraisal," *Oxford Review of Education*, Volume 30, Number 4, 2004, pp. 451-465.
- Cancer Prevention Resource Center "CPRC Measures,"
<http://www.uri.edu/research/cprc/measures.htm> accessed on 4/16/07.
- Carroll, J. M., "Five reasons for scenario-based design," *Proceedings of the HICSS-32, Hawaii*, 1999, pp. 11.
- Carter, K., "Do teachers understand the principles for writing tests?" *Journal of Teacher Education*, Volume 35, Number 6, 1984, pp. 57-60.
- Chi, M. T. H., Siler, S. A., and Jeong, H., "Can Tutors Monitor Students' Understanding Accurately?" *Cognition and Instruction*, Volume 22, Number 3, 2004, pp. 363-387.
- Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., and Hausmann, R. G., "Learning from Human Tutoring," *Cognitive Science*, Volume 25, Number 4, 2001, pp. 471-533.
- Chin, J. P., Diehl, V. A., and Norman, K. L., "Development of an Instrument Measuring User Satisfaction of the Human-Computer Interface," *Proceedings of the CHI '98, New York*, 1988.
- Crooks, T. J., "The Impact of Classroom Evaluation on Students," *Review of Educational Research*, Volume 58, Number 4, 1988, pp. 438-81.
- Darling-Hammond, L., and Youngs, P., "Defining "Highly Qualified Teachers": What Does "Scientifically-Based Research" Actually Tell Us?" *Educational Researcher*, Volume 31, Number 9, 2002,

- Deci, E. L., Koestner, R., and Ryan, R. M., "Extrinsic Rewards and Intrinsic Motivation in Education: Reconsidered Once Again," *Review of Educational Research*, Volume 71, Number 1, 2001, pp. 1-27.
- DeSanctis, G., and Poole, M. S., "Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory," *Organization Science*, Volume 5, Number 2, 1994, pp. 121-147.
- Ellett, C. D., and Teddlie, C., "Teacher Evaluation, Teacher Effectiveness and School Effectiveness: Perspectives from the USA," *Journal of Personnel Evaluation in Education*, Volume 17, Number 1, 2003, pp. 101-128.
- Fielding, R., "Architectural Styles and the Design of Network-based Software Architectures," University of California at Irvine, 2000.
- Fisher, T. H. "Part 2: A Review and Analysis of the Tennessee Value-Added Assessment System," (State of Tennessee, Comptroller of the Treasury, Office of Education Accountability), 1996, pp. 58.
- Fishman, B., Best, S., Foster, J., and Marx, R., "Fostering Teacher Learning in Systemic Reform: A Design Proposal for Developing Professional Development," *Proceedings of the NARST 2000*, New Orleans, LA, 2000, pp. 16.
- Fishman, B., Marx, R. W., Best, S., and Tal, R. T., "Linking Teacher and Student Learning to Improve Professional Development in Systemic Reform," *Teaching and Teacher Education*, Volume 19, Number 6, 2003, pp. 643-658.
- Fleming, M., and Chambers, B., "Teacher-made Tests: Windows on the Classroom," In *Testing in the Schools: New Directions for Testing and Measurement*, W. E. Hathaway (Ed.), Jossey-Bass, San Francisco, 1983, pp. 29-38.
- Fogg, B. J., "Persuasive Computers: Perspectives and Research Directions," *Proceedings of the CHI'98*, Los Angeles, CA, 1998, pp. 225-32.
- Fogg, B. J., "Introduction to the Special Section on Persuasive Technologies," *Communications of the ACM*, Volume 42, Number 5, 1999, pp. 27-29.
- Fogg, B. J., and Nass, C., "How Users Reciprocate to Computers: An experiment that demonstrates behavior change," *Proceedings of the CHI'97*, Atlanta, GA, 1997,
- Garson, D. "Case Studies." <http://www2.chass.ncsu.edu/garson/pa765/cases.htm> accessed on 4/11/07.
- Goodhue, D. L., and Thompson, R. L., "Task-Technology Fit and Individual Performance," *MIS Quarterly*, Volume 19, Number 2, 1995, pp. 213-236.

- Green, K. E., "Multiple-Choice Item Difficulty: The Effects of Language and Distracter Set Similarity," Proceedings of the 67th Annual Meeting of the American Educational Research Association, Montreal, Quebec, CAN, 1983.
- Haladyna, T. M., Writing Multiple-Choice Items, Computer Adaptive Technologies, Inc., Evanston, IL, 2001.
- Haladyna, T. M., and Downing, S. M., "A Taxonomy of Multiple-Choice Item-Writing Rules," *Applied Measurement in Education*, Volume 2, Number 1, 1989a, pp. 37-50.
- Haladyna, T. M., and Downing, S. M., "Validity of a Taxonomy of Multiple-Choice Item-Writing Rules," *Applied Measurement in Education*, Volume 2, Number 1, 1989b, pp. 51-78.
- Haladyna, T. M., Downing, S. M., and Rodriguez, M. C., "A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment," *Applied Measurement in Education*, Volume 15, Number 3, 2002, pp. 309-334.
- Hambleton, R. K., and Rogers, H. J., "Item Bias Review," *Practical Assessment, Research & Evaluation*, Volume 4, Number 6, 1995.
- Hambleton, R. K., Swaminathan, H., and Rogers, H. J., *Fundamentals of Item Response Theory*, Sage Publications, 1991.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992) Force Concept Inventory, *Physics Teacher*, 30, pp. 141-158.
- Hill, D. "He's Got Your Number," *Teacher Magazine*, Volume 11, Number 8, May, 1st 2000, pp. 42-47.
- Hirst, G.; DiMarco, C.; Hovy, E.; and Parsons, K., "Authoring and Generating Health-Education Documents That Are Tailored to the Needs of the Individual Patient," *Proceedings of the User Modeling: Proceedings of the Sixth International Conference, UM97, 1997*, pp. 107-118.
- Iszák, A., and Sherin, M. G., "Exploring the Use of New Representations as a Resource for Teacher Learning," *School Science and Mathematics*, Volume 103, Number 1, 2003, pp. 18-27.
- Janis, I. L., and Mann, L., *Decision Making: A Psychological Analysis of Conflict, Choice and Commitment*, Free Press (MacMillan), 1977.
- Koedinger, K. R., "Intelligent Cognitive Tutors as Modeling Tool and Instructional Model," *Proceedings of the NCTM Standards 2000 Technology Conference*, Arlington, VA, 1998, pp. 1-13.

- Kubitskey, B., and Fishman, B., "Untangling the Relationship(s) Between Professional Development, Enactment, Student Learning and Teacher Learning Through Multiple Case Studies," Proceedings of the American Educational Research Association, Montreal, Canada, 2005, pp. 29.
- Kupermintz, H., "Teacher Effects as a Measure of Teacher Effectiveness: Construct Validity Considerations in TVAAS (Tennessee Value-Added Assessment System)," CSE Technical Report 563, CRESST/University of Colorado, Boulder, 2002.
- Kyriacou, C., "Teacher Stress: directions for future research," Educational Review, Volume 53, Number 1, 2001, pp. 27-35.
- Leidner, D. E., and Jarvenpaa, S. L., "The Use of Information Technology to Enhance Management School Education: A Theoretical View," MIS Quarterly, Volume 19, Number 3, 1995, pp. 265-291.
- Lenth, R. V., "Some Practical Guidelines for Effective Sample-Size Determination," The American Statistician, Volume 55, 2001, pp. 187-193.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., and Hamilton, L. S., "Evaluating Value-Added Models for Teacher Accountability," MG-158, RAND Corporation, 2003.
- McGrenere, J. "An Evaluation of a Multiple Interface Design Solution for Bloated Software," Proceedings of CHI 2002, ACM CHI Letters 4(1), pp. 163-170.
- McKeachie, W. J., *McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers*, Houghton Mifflin, Boston, 2002.
- McKeown, M., and Beck, I. L., "Transforming Knowledge into Professional Development Resources: Six Teachers Implement a Model of Teaching for Understanding Text," The Elementary School Journal, Volume 104, Number 5, 2004, pp. 391-408.
- Minstrell, J., "Facets of students' knowledge and relevant instruction," In *Research in physics learning: Theoretical issues and empirical studies*, R. Druit, F. Goldberg and H. Niedderer (Ed.), University of Kiel, Institute for Science Education, Kiel, 1992, pp. 110-128.
- Nass, C., Moon, Y., Fogg, B. J., Reeves, B., and Dryer, C., "Can Computer Personalities Be Human Personalities?" Proceedings of the CHI'95, Denver, CO, 1995, pp. 228-9.
- Nass, C., Steuer, J., and Tauber, E., "Computers are Social Actors," Proceedings of the CHI'94, Boston, MA, 1994, pp. 72-78.

- National Council of Teachers of Mathematics, "Principles and standards for school mathematics," 2000.
- National Research Council, National Science Education Standards, National Academy Press, Washington, D.C., 1996.
- Newman, L. S., and Taube, K. T., "The Accuracy and Use of Item Difficulty Calibrations Estimated from Judges' Ratings of Item Difficulty," Proceedings of the Annual Meeting of the American Educational Research Association, New York, NY, USA, 1996,
- O'Loughlin, M., "Rethinking Science Education: Beyond Piagetian Constructivism Toward a Sociocultural Model of Teaching and Learning," *Journal of Research in Science Teaching*, Volume 29, Number 8, 1992, pp. 791-820.
- Orlikowski, W., "The Duality of Technology: Rethinking the Concept of Technology in Organizations," *Organization Science*, Volume 3, Number 3, 1992, pp. 398-427.
- Parshall, C. G., Kromrey, J. D., Chason, W. M., and Yi, Q., "Evaluation of Parameter Estimation under Modified IRT Models and Small Samples," Proceedings of the Annual Meeting of the Psychometric Society, Gatlinburg, TN, 1997, pp. 2-64.
- Prestwood, J. S., and Weiss, D. J., "Accuracy of Perceived Test-Item Difficulties," 77-3, Office of Naval Research, Arlington, VA. Personnel and Training Research Programs Office, 1977.
- Prochaska, J. M., Prochaska, J. O., and Levesque, D. A., "A Transtheoretical Approach to Changing Organizations," *Administration and Policy in Mental Health*, Volume 28, Number 4, 2001a, pp. 247-261.
- Prochaska, J. O., and Velicer, W. F., "The Transtheoretical Model of Health Behavior Change," *American Journal of Health Promotion*, Volume 12, Number 1, 1997, pp. 38-48.
- Prochaska, J. O., Velicer, W. F., Fava, J. L., Rossi, J. S., and Tosh, J. Y., "Evaluating a population-based recruitment approach and a stage-based expert system intervention for smoking cessation," *Addictive Behaviors*, Volume 26, 2001b, pp. 583-602.
- Prochaska, J. O., Velicer, W. F., Rossi, J. S., Goldstein, M. G., Marcus, B., Rakowski, W., Fiore, C., Harlow, L. L., Redding, C. A., Rosenbloom, D., and Rossi, S. R., "Stages of change and decisional balance for 12 problem behaviors," *Health Psychology*, Volume 13, 1994, pp. 39-46.
- Richardson, E. "Rules for Preparing Educators," <http://www.alsde.edu/>, accessed 3/15/2006, 2003a.

- Richardson, V., "The role of attitudes and beliefs in learning to teach," In *Handbook of Research on Teacher Education*, J. Sikula, T. Buttery and E. Guyton (Ed.), Simon & Shuster Macmillan, New York, 1996, pp. 102-119.
- Richardson, V., "The Dilemmas of Professional Development," *Phi Delta Kappan*, Volume 84, Number 5, 2003b, pp. 401-406.
- Rowe, K. J., "The Importance of Teacher Quality as a Key Determinant of Students' Experiences and Outcomes of Schooling," *Proceedings of the ACER Research Conference*, Melbourne, Australia, 2003, pp. 1-51.
- Rowe, M. B., "Wait time and rewards as instructional variables, their influence on language, logic, and fate control," *Journal of Research in Science Teaching*, Volume 11, Number 2, 1974, pp. 81-94.
- Sadler, R. D., "Formative Assessment and the Design of Instructional Systems," *Instructional Science*, Volume 18, Number 2, 1989, pp. 119-44.
- Sanders, W., and Horn, S. P., "The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment," *Journal of Personnel Evaluation in Education*, Volume 8, 1994, pp. 299-311.
- Saunders, L., "A Brief History of Educational 'Value Added': How did we get to where we are?" *School Effectiveness and School Improvement*, Volume 10, Number 2, 1999, pp. 233-256.
- Sherin, M. G., "When Teaching Becomes Learning," *Cognition and Instruction*, Volume 20, Number 2, 2002, pp. 119-150.
- Shulman, L. S., "Those who understand: Knowledge growth in teaching," *Educational Researcher*, Volume 15, Number 2, 1986, pp. 4-14.
- Singley, M. K., and Bennett, R. E., "Item generation and beyond: Applications of schema theory to mathematics assessment," In *Item generation for test development*, S. Irvine and P. Kyllonen (Ed.), Earlbaum, Hillsdale, NJ, 2002.
- Smith, W. R., "Evidence for the Effectiveness of Techniques to Change Physician Behavior," *CHEST*, Volume 118, Number 2, 2000, pp. 8S-17S.
- Stiggins, R. J., "The Unfulfilled Promise of Classroom Assessment," *Educational Measurement: Issues and Practice*, Volume 20, Number 3, 2001, pp. 5-15.
- Stiggins, R. J., Frisbie, D. A., and Griswold, P. A., "Inside High School Grading Practices: Building a Research Agenda," *Educational Measurement: Issues and Practice*, Volume 8, Number 2, 1989, pp. 5-14.
- Stroup, W. W. "Assessment of the Statistical Methodology Used in the Tennessee Value-Added Assessment System," (Tennessee Value Added Assessment Center), 1995.

- Thissen-Roe, A., Hunt, E., and Minstrell, J., "The DIAGNOSER project: Combining assessment and learning," *Behavior Research Methods, Instruments, & Computers*, Volume 36, Number 2, 2004, pp. 234-240.
- Thoennessen, M., Kashy, E., Tsai, Y., and Davis, N. E., "Impact of Asynchronous Learning Networks in Large Lecture Classes," *Group Decision and Negotiation*, Volume 8, 1999, pp. 371-384.
- VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. B., "Why Do Only Some Events Cause Learning During Human Tutoring?" *Cognition and Instruction*, Volume 21, Number 3, 2003, pp. 209-249.
- Velicer, W. F., DiClemente, C. C., Prochaska, J. O., and Brandenburg, N., "A decisional balance measure for assessing and predicting smoking status," *Journal of Personal and Social Psychology*, Volume 48, 1985, pp. 1279-1289.
- Velicer, W. F., and Prochaska, J. O., "An Expert System Intervention for Smoking Cessation," *Patient Education and Counseling*, Volume 36, 1999, pp. 119-129.
- Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D., "User Acceptance of Information Technology: Toward a Unified View," *MIS Quarterly*, Volume 27, Number 3, 2003, pp. 425-478.
- Walter, I., Nutley, S., and Davies, H., "Developing a Taxonomy of Interventions Used to Increase the Impact of Research," *Economic & Social Research Council UK Centre for Evidence Based Policy & Practice*, February 2003.
- Whitmer, S. P., "A descriptive multi-method study of teacher judgement during the marking process," *Research Series No. 122*, Michigan State University, Institute for Research on Teaching, 1983.
- Yamagata-Lynch, L. C., "How a technology professional development program fits into teachers' work life," *Teaching and Teacher Education*, Volume 19, 2003, pp. 591-607.